
Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

# Modelo de predicción sobre deserción académica en educación superior de la población vulnerable

Presentado por:

María del Pilar Giraldo Giraldo, Jaime Enrique Hernández Poveda

Bogotá, D.C., Colombia, 17 de junio de 2023


Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

# Modelo de predicción sobre deserción académica en educación superior de la población vulnerable

Presentado por:

María del Pilar Giraldo Giraldo, Jaime Enrique Hernández Poveda

Bajo la dirección de:

Jeison Orlando Pinilla Alzate

Bogotá, D.C., Colombia, 17 de junio de 2023

**Tabla de contenido**

## Agradecimientos

Agradecemos a nuestro director de proyecto Jeison Pinilla Alzate por su oportuna asesoría, guía y retroalimentación en el transcurso de la maestría, así como a los demás docentes, que con sus enseñanzas aportaron a la construcción de este proyecto empresarial.

Así mismo, agradecemos al Ministerio de Educación Nacional y al ICETEX por suministrarnos la información requerida de forma oportuna y veraz. Por otra parte, al Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia por apoyar la educación de posgrado en el área de tecnologías de la información en nuestro país.

*Maria del Pilar Giraldo Giraldo*

*Jaime Enrique Hernández Poveda*

## Dedicatoria

Este proyecto es un homenaje para las personas que a través de la vida me han apoyado en mis sueños, que día a día me motivan a ser mejor persona, con valores y deseos de ayudar a los demás; estas personas son mis padres Maria del Carmen y Jaime Enrique que nunca se han cansado de incentivarme a estudiar. A mi hermano Iván que ha sido como un profesor y un ejemplo para tomar buenas decisiones en mi vida. A mi pareja Jury, que con su apoyo y sacrificio no me dejó desfallecer a lo largo de esta maestría. Por último y más importante, a Dios que me permite todos los días aprender, soñar, disfrutar y ser feliz en este camino llamado vida.

*Jaime Enrique.*

Dedico este proyecto a mi esposo, mi hijo y mi padre, por ser los principales agentes motivadores en mi vida. Agradezco a Dios por su presencia en mi vida, a los viejos amigos, así como a los más recientes, que, con su apoyo constante, una palabra de aliento y una frase cálida, fueron de gran soporte para cumplir esta nueva meta.

*María del Pilar.*

**Declaración de originalidad y autonomía.**

Declaramos bajo la gravedad del juramento, que hemos escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi(nuestra) propia cuenta y que, por lo tanto, su contenido es original.

Declaramos que hemos indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

María del Pilar Giraldo Giraldo

Jaime Enrique Hernández Poveda

Firmado en Bogotá, D.C. el 17 de junio de 2023

**Declaración de exoneración de responsabilidad**

Declaramos que la responsabilidad intelectual del presente trabajo es exclusivamente de sus autores. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

María del Pilar Giraldo Giraldo

Jaime Enrique Hernández Poveda

Firmado en Bogotá, D.C. el 17 de junio de 2023

## Lista de tablas

## Lista de Figuras

## Lista de Apéndices

# Glosario

Deserción Escolar: “Abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno”(Ministerio de Educación Nacional de Colombia, 2017a).

IES: Institución de Educación Superior

Población Vulnerable: Grupo de personas que se encuentran en estado de desprotección o incapacidad frente a una amenaza a su condición psicológica, física y mental, entre otras. En el ámbito educativo este término hace referencia al grupo poblacional excluido tradicionalmente del sistema educativo por sus particularidades o por razones socioeconómicas. (Ministerio de Educación Nacional de Colombia, 2017b)

## Resumen Ejecutivo

Para el proyecto empresarial de la maestría en Business Analytics se creará un Modelo de predicción sobre deserción académica en educación superior de la población vulnerable con altos resultados en las pruebas saber 11 del país, la población a analizar son los estudiantes en condición de desertados y en condición de graduados del programa de gobierno llamado Ser Pilo Paga, que ingresaron a Instituciones de Educación Superior acreditadas en alta calidad.

La primera actividad por realizar es definir el número de personas condonadas o graduadas, así como inactivas o desertadas del total de la data, posterior a ello se realizará la validación de las variables socioeconómicas y académicas más significativas frente a la deserción académica para lograr la identificación de posibles nuevos desertores.

Teniendo en cuenta lo anterior, gracias a los conocimientos adquiridos, se explorará las diferentes técnicas, así como determinar el modelo predictivo a usar, por medio de la comprobación y mayor porcentaje de efectividad una vez se corran los diferentes escenarios que incorporen las variables de la data.

Posteriormente, se realizará una evaluación de los escenarios seleccionados, con el fin de seleccionar el más efectivo, realizar la predicción de la deserción frente a los beneficiarios activos académicamente. Para finalizar se realiza el backtesting con los beneficiarios que cambiaron su condición de estudiantes a condonados graduados o inactivos desertores.

**Palabras clave**

Educación superior, Modelo estadístico, Población vulnerable, Deserción, Predicción.

## Abstract

For the business project of the master's degree in Business Analytics, a prediction model will be created on academic desertion in higher education of the vulnerable population with high results in the saber 11 tests of the country, the population to be analyzed are the students in deserted condition and in condition of graduates of the government program called Ser Pilo Paga, who entered Higher Education Institutions accredited in high quality.

The first activity to be carried out is to define the number of people condoned or graduated, as well as inactive or deserted from the total data, after which the validation of the most significant socioeconomic and academic variables will be carried out against academic desertion to achieve identification. of potential new dropouts.

Taking into account the above, thanks to the knowledge acquired, the different techniques will be explored, as well as determining the predictive model to use, through verification and a higher percentage of effectiveness once the different scenarios that incorporate the variables of the analysis are run. data.

Subsequently, an evaluation of the selected scenarios will be carried out, in order to select the most effective, to make the prediction of the dropout against the academically active beneficiaries. Finally, backtesting is carried out with the beneficiaries who changed their status from students to condoned graduates or inactive dropouts.

**Keywords**

Higher education, Statistical model, Vulnerable population, Desertion, Prediction.

# 1. Introducción

El Gobierno Nacional Colombiano debe garantizar el acceso y permanencia a la educación de la población colombiana de acuerdo con lo señalado en el artículo 44, de la Constitución Política de la República de Colombia (1991).

La población vulnerable en Colombia, no cuenta con las herramientas suficientes que le permita tomar una decisión frente a la elección de un programa académico de formación profesional y a partir de los conocimientos adquiridos en su proceso de formación básica y media, lo cual conlleva con el tiempo al desinterés en la formación académica (Betancourth Sánchez & Cuesta, 2016), generación de profesionales no felices, poco interés en continuar estudios de posgrados, así como falta de motivación para continuar con el crecimiento académico y profesional, así que la suma de todos estos aspectos, entre otros, pueden llevar a una posible deserción académica (Martínez Ipuz et al., 2009).

Como parte de la solución a esta problemática, el Gobierno Nacional ha dispuesto recursos finitos (Mora Cortés, 2016) para garantizar el acceso a la educación superior con la creación de programas y políticas públicas, entre ellos a través del otorgamiento de créditos condonables; sin embargo, no se tiene la seguridad del aprovechamiento del 100% de estos recursos de manera eficiente, motivo por el cual se podría incurrir en un posible detrimento patrimonial.

Así mismo las Instituciones de Educación Superior otorgan cupos en cada semestre académico, con el objetivo de generar nuevos profesionales, no obstante, si los nuevos posibles estudiantes no están bien orientados en la selección del programa a cursar, existe una alta probabilidad de que en el desarrollo del pregrado se presente una posible deserción académica y con ello la pérdida de estos cupos.

Conforme con lo expuesto, se espera diseñar un modelo de predicción sobre deserción en educación superior de la población vulnerable, que según el Ministerio de Educación Nacional se define como: "Grupo de personas que se encuentran en estado de desprotección o incapacidad frente a una amenaza a su condición psicológica, física y mental, entre otras. En el ámbito educativo este término hace referencia al grupo poblacional excluido tradicionalmente del sistema educativo por sus particularidades o por razones socioeconómicas" (Ministerio de Educación Nacional de Colombia, 2017b), correspondiente al programa ser pilo paga, con el fin de presentar un estudio a la problemática existente. Es de señalar que la población del proyecto fue definida por el Ministerio de Educación Nacional como población vulnerable, teniendo en cuenta su condición socioeconómica, definida en el puntaje Sisbén II y que, a su vez, lograron altos puntajes en las pruebas Saber 11.

Para desarrollar este proyecto empresarial, se analizará y explorará las técnicas de análisis de datos, probabilidad y estadística que se requieran tales como (visualización de datos a través de Power BI, analítica descriptiva, analítica predictiva, bases de datos, Analitycs life cycle managment, Análisis de riesgo, Análisis de componentes principales

(ACP) y Análisis de correspondencias Simples (ACS), entre otras). La matriz de riesgos, así como el cronograma del proyecto empresarial se encuentra en los apéndices A y B respectivamente.

Así mismo se usará la base de datos del programa ser pilo paga (en condición de desistidos y graduados al corte de 30 de junio de 2022), de igual forma se podrá apoyar en bases de datos públicas del ICFES, SISBEN entre otras, en el caso de necesitarlas.

Cabe resaltar que en el desarrollo del proyecto empresarial, el Ministerio de Educación Nacional suministró nueva información asociada al núcleo familiar de los beneficiarios del programa ser pilo paga, por lo anterior, se inició el proyecto con una cantidad menor de variables y al final aumentó la cantidad de variables a 27; así mismo se definió la eliminación de la variable "Numero Semestres Adjudicados" dado que no cumplía con la calidad requerida para análisis y experimentación, motivo por el cual se excluyó de los análisis exploratorios. Finalmente se registró la información más relevante en el desarrollo del modelo.

Es importante de igual forma tener en cuenta que el programa ser Pilo Paga se materializó en el año 2015, bajo el Gobierno del expresidente Juan Manuel Santos, con el fin de generar alrededor de 40 mil créditos condonables, bajo el cumplimiento del requisito: graduación del programa financiado; como estrategias establecidas por el gobierno nacional para fomentar el acceso a la educación superior de la población menos favorecida y con grandes resultados académicos en su prueba de estado.

# 2. Objetivos

## General

Utilizar herramientas analíticas que permitan medir los procesos de acceso, permanencia y graduación en la educación superior del país, a través de un modelo de predicción de la deserción de la población del programa ser pilo paga, en la educación superior, analizando las variables socioeconómicas y académicas de los estudiantes.

## Específicos

Identificar las variables que estadísticamente infieren en la deserción académica de los jóvenes del programa ser Pilo Paga en sus 4 versiones, siendo estos perteneciente a un segmento de la población vulnerable con resultados académicos sobresalientes.

Estimar un modelo que permita la utilización de variables académicas y socioeconómicas, de tal forma que explique la deserción académica o graduación de la población realizando un proceso de clasificación.

Calcular las medidas de validación y verificación del modelo analítico a usar de acuerdo con la calidad, tipo y condición de la información registrada en la data, para así lograr la mayor efectividad en la predicción.

Aplicar el modelo seleccionado en la población del programa Ser Pilo Paga que continua en curso sus estudios para determinar el escenario de posibles desertores y graduados.

Realizar un Backtesting al modelo con nuevos datos suministrados, comparando la situación real de los estudiantes activos que se hayan graduado o desertado de su programa académico versus la predicción inicial arrojada por el modelo.

# 3. Alcance

Por medio de la analítica de datos aplicada a la base de datos del programa Ser Pilo Paga y en especial a la población de estudiantes condonados o graduados y estudiantes inactivos o desertados, se busca identificar las variables que tengan correlación y significancia con la problemática de la deserción académica en la educación superior de la población vulnerable en Colombia, quienes no cuentan con el poder adquisitivo para acceder a la formación profesional en una Institución de Educación Superior acreditada en alta calidad, de esta forma, se busca crear un modelo de alta probabilidad de ocurrencia o no, de la deserción académica de los estudiantes en comento, que a un momento específico se encuentren graduados o por el contrario hayan desertado del programa académico.

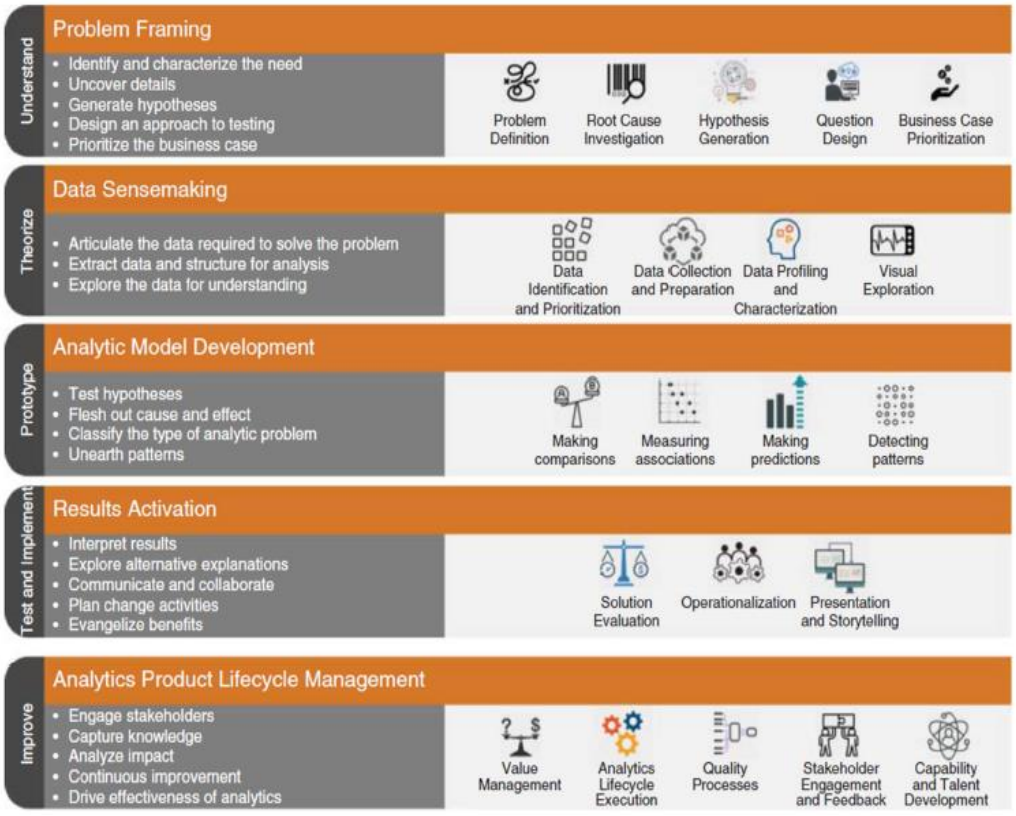
Con este modelo predictivo, se podrá identificar las diferentes variables que inciden en la deserción académica.

# 4. Metodología

Antes de realizar un modelo de predicción sobre deserción académica de población vulnerable del programa ser pilo paga, vamos a revisar diferentes aspectos que se deben realizar con el ánimo de entender el negocio, conocer al detalle la información que se ha recibido, para poder entender claramente el problema a resolver y establecer una ruta clara que permita alcanzar el cumplimiento del objetivo general, para ello se usará la metodología descrita en la Figura 1, aclarando que la metodología ágil se encuentra inmersa en el desarrollo del proyecto.

**Figura 1**

*Metodología*




*Nota:* Metodología para desarrollar proyecto empresarial. Fuente: Diapositivas Analytics Life Cycle Management, (2023)

Partiendo entonces de esta metodología, el primer paso a seguir es la definición del problema.

### 4.1. Definición del problema - Problema del negocio:

La misión del ICETEX es poder generar Acceso, permanencia y graduación en la educación superior. Dentro de la oferta entregada por el ICETEX se encuentran los Fondos en Administración que corresponden a recursos de terceros entregados a esta entidad para fomentar el acceso y permanencia en la educación superior, siendo este el caso del programa ser Pilo Paga, con el cual se realizó otorgamiento de créditos condonables a la población más vulnerable y con mejores resultados en las pruebas Saber 11. En esta población objetivo se logró identificar que algunos de los jóvenes beneficiarios de este programa, no culminaron sus estudios, por lo anterior, se presentó la deserción académica que impidió la graduación de estos.

Una vez entendido el problema del negocio y establecidos los objetivos generales y específicos, se plasmaron las siguientes historias de usuario:

**Tabla 1**

*Historias de usuario*

| Historia de usuario | Prioridad (Alta, media, baja) | Criterios de aceptación |
|---|---|---|
| 1. Soy el Ministerio de Educación Nacional, deseo saber cuáles son las variables que impactan en la deserción o graduación académica de un estudiante de población vulnerable y con resultados académicos sobresalientes | 1. [Alta] | 1. Listado de variables estadísticamente significativas de la población vulnerable ser pilo paga que influyen en la deserción o graduación. |
| 2. Como investigadores del problema de negocio, necesitamos determinar frente a los modelos predictivos existentes (supervisados, no supervisados, redes neuronales y no tradicionales) cual desarrolla una mejor predicción. | 2. [Medio] | 1. Obtener un modelo que estadísticamente refleja una mejor efectividad basado en su R^2, AUC, curva Roc o indicador que corresponda. |
| 3. Como Ministerio de Educación Nacional necesito conocer el comportamiento probable sobre deserción o graduación de los beneficiarios de Ser Pilo Paga que se encuentran con estudios en curso. | 3. [Alta] | 1. Listado de estudiantes activos académicamente con definición de su posible graduación o deserción |
| 4. Como lideres de analítica es necesario realizar la evaluación del modelo, es decir, verificar la predicción modelada contra la situación real de los estudiantes que se encontraban activos | 4. [Medio] | 1. Un indicador porcentual calculado de la siguiente manera = número de aciertos del modelo / No. de beneficiarios reales graduados o desertados |

*Nota*, Historia de Usuario con prioridad y criterios de aceptación. Fuente: Elaboración Propia, (2023)

### 4.2.Problema Analítico:

Ahora bien a través de la utilización de herramientas analíticas, se pretende generar estrategias que permitan fortalecer los procesos de acceso, permanencia y graduación en la educación superior del país, para ello se desarrollará un modelo de predicción de la deserción de la educación superior de la población del programa ser pilo paga; con el cual se realizaran análisis de correlación de las diferentes variables cuantitativas, que puedan explicar la situación presentada con los estudiantes y la decisión del retiro del programa académico en la Institución de Educación Superior

### 4.3.Investigación de causa raíz:

En el proceso de investigación de la causa raíz se desarrollan diferentes actividades, entre ellas se encuentra:

1. Consultar los textos bibliográficos que contengan estudios previos de las causas de deserción de la población vulnerable.
2. Hacer ejercicios de correlación de las variables con respecto a la deserción o no de la población estudiada.

A partir de estas actividades se inicia el proceso de creación o generación de hipótesis.

### 4.4.Generación de hipótesis:

- Creemos que las fortalezas académicas de los estudiantes disminuyen la deserción, se logra si los estudiantes con mejores resultados en las pruebas Saber 11 culminan sus estudios exitosamente.
- Creemos que la procedencia de los estudiantes ubicados en zonas rurales no permitirá el buen desempeño académico en la universidad, se logra si los jóvenes no culminan el programa y se convierten en desertores.
- Creemos que la condición económica y social (puntaje Sisbén II) de la población estudiantil, permitirá explicar la graduación o no del programa académico, se logra si la población con menor puntaje presenta dificultad en culminar sus estudios de pregrado.

- Creemos que los subsidios de sostenimiento asignados por parte del Ministerio de Educación Nacional influyen en la deserción académica de los estudiantes, se logra si la población que tiene mayor subsidio asignado se gradúa del programa cursado.

**4.5.Enfoque:**

Dentro del proyecto a realizar, se define en un principio utilizar variables que posiblemente tengan correlación o significancia con la problemática de la deserción académica en la educación superior de la población vulnerable en Colombia, quienes no cuentan con el poder adquisitivo para acceder a una Institución de Educación Superior acreditada en alta calidad, por lo anterior, se busca crear un modelo de alta probabilidad de ocurrencia o no de la deserción académica de la muestra de los estudiantes en comento, que en un momento específico se encuentren graduados o por el contrario que hayan desertado del programa académico.

De dicha población se segmentaron los estudiantes que a la fecha de corte 30 de junio de 2022, se encontraban graduados, desertados o que continúan en época de estudio, ahora bien, se utilizará las variables socioeconómicas, demográficas, académicas y otras que se tenían en el momento en que se adjudicaron los créditos.

**4.6.Priorización de los casos de negocio:**

Dentro de las actividades a realizar se definen:

1. Solicitar la base de datos del programa ser pilo paga al Ministerio de Educación Nacional e ICETEX para realizar el proyecto analítico.

2. Revisión de los datos, estableciendo las prioridades y definición del proyecto analítico.

3. Acuerdo final de las partes del proyecto analítico a realizar, en donde se define el alcance, la metodología, cronograma y el modelo que se utilizará para la predicción de la deserción académica.

4. Presentación final del proyecto analítico empresarial.

### 4.7.Sentido de datos - Identificación y Priorización

Las actividades por realizar:

1. Identificar la información requerida para el proyecto analítico, así como la entregada por la organización, validar los conceptos básicos, e indagar sobre las dudas que se puedan generar.

2. Identificar las variables faltantes y sus fuentes de recolección (si aplica).

3. Realizar un análisis de la calidad de los datos.

4. Realizar un glosario de la información a utilizar donde explique las características de esta.

### 4.8.Recopilación y preparación de datos

Las actividades por realizar:

1. Seleccionar las fuentes de información, para este caso se nos entregó una base oficial anonimizada del programa Ser Pilo Paga en sus 4 versiones, la cual ya incluía la información de ICETEX en el momento de adjudicación (Fuentes persistentes) y la información externa del SISBEN e ICFES (Fuentes alternas).

2. Depurar, validar y limpiar las bases de información acorde a la calidad de datos deseada y concluyendo la pertinencia de la utilización de las variables proyectadas para el estudio.

3. Generar conclusiones de los hallazgos y descripciones de la revisión de los datos y fuentes con el cual podamos definir la viabilidad del proyecto.

4. Analizar y determinar con los resultados obtenidos, la(s) herramienta(s) analítica requerida para el desarrollo del proyecto que nos brinde las mejores posibilidades para el modelaje analítico.

### 4.9.Perfilado y Caracterización de Datos

Las actividades por realizar:

1. Realizar el reconocimiento de las variables que se encuentran en la data, así como el entendimiento de estas. De esta forma iniciar con la definición del tipo de variables que se manipularan (variables continuas y categóricas).

2. Descripción estadística de las variables revisando el comportamiento de los datos por medio de medidas de tendencia, dispersión, forma, extremos, entre otros.

3. Realizar las primeras exploraciones de las relaciones que pueden llegar a tener las variables independientes con la variable dependiente (Deserción), siempre teniendo en cuenta los supuestos estadísticos con respecto a la significancia de las variables.

4. Con los resultados obtenidos y al encontrar temas atípicos o inusuales, generar un proceso adicional de calidad de datos para confirmar el comportamiento de las variables.

## 4.10. Exploración Visual

Se realiza un Dashboard en Power Bi con la data recibida del programa ser pilo paga, que ilustra el comportamiento inicial de los beneficiarios, en una exploración general en la que se puede ver la versión del programa, así como otros factores que pueden ser relevantes.

En esta visualización, se logró plasmar los primeros análisis exploratorios de las variables, en los cuales se identifican los siguientes aspectos:

Con respecto al género, se pudo concluir que el 57,12% corresponde a Masculino y 42,88% Femenino.

El promedio de las pruebas Saber 11 de la totalidad de estudiantes es 347,5 puntos.

La mayoría de los estudiantes se concentra en los estratos 2 y 1 con total de 16.903 y 16.110 beneficiarios respectivamente.

5.283 estudiantes corresponden a la ciudad de Bogotá siendo el municipio más representativo del top 10, y el menor el municipio de Villavicencio con 648 estudiantes.

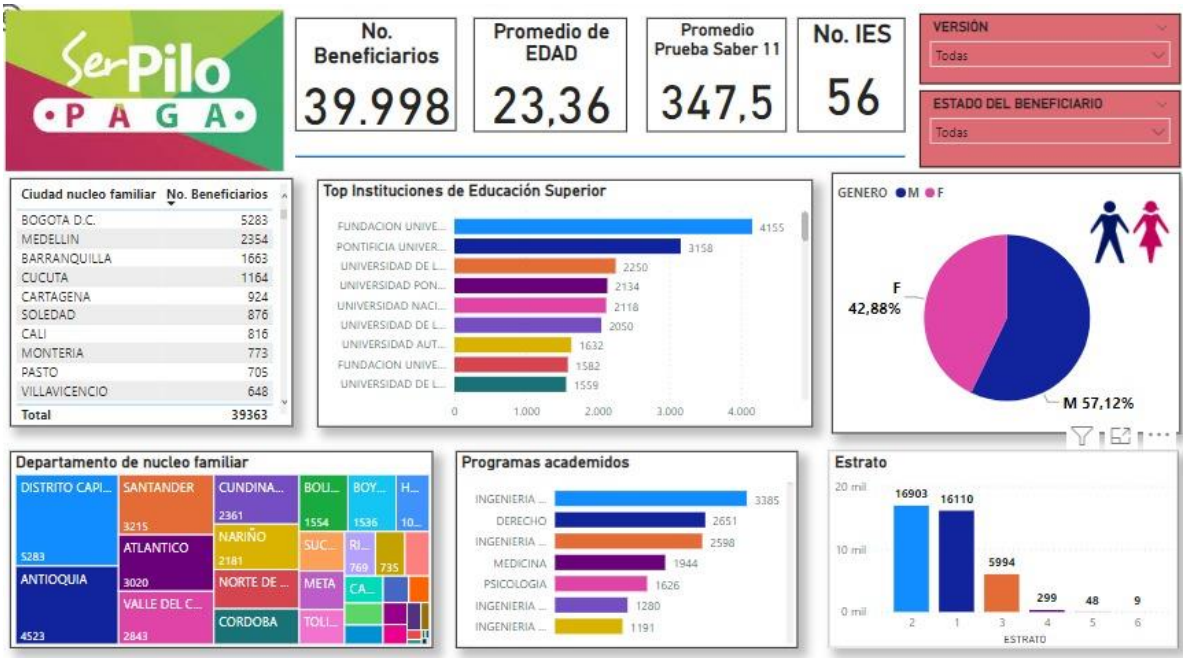
Participaron 56 Instituciones de Educación Superior que se encuentran calificadas con registro de alta calidad.

Del Top de las 9 Instituciones de Educación Superior lidera la Fundación Universidad del Norte con 4.155 estudiantes.

El programada con mayor número de estudiantes es Ingeniería Civil con un total 3.385 jóvenes.

**Figura 2**

*Tablero visualización Data*




*Nota.* Muestra la información general de los datos del programa ser pilo paga a junio de 2022. Fuente: Elaboración Propia, (2022)

## 5. Descripción de la Situación organizacional donde se realizará el proyecto

La aplicación del proyecto empresarial se fundamenta en el programa Ser Pilo Paga, este programa fue creado por el Gobierno de turno de Juan Manuel Santos, en el cual se diseñó una estrategia desde el Ministerio de Educación Nacional, fundamentada en disminuir las brechas existentes para la población más vulnerable del país, teniendo en cuenta que este grupo de personas no puede acceder a la educación superior de alta calidad dado que no tienen los recursos para costear los costos de matrícula en una universidad pública ni mucho menos en una privada, es por ello, que el Gobierno diseñó el otorgamiento de créditos condonables a estudiantes del grado 11 que tuvieran condición económica vulnerable y a su vez tuvieran excelentes resultados en pruebas saber 11 (Mora Cortés, 2016).

Durante los 4 años de Gobierno, la meta fue adjudicar 40.000 créditos condonables para estudios de educación superior en las Universidades certificadas en alta calidad o con programas de alta calidad, anualmente se entregarían 10.000 cupos a las personas que tuvieran un puntaje Sisbén inferior, pero a su vez tuvieran los puntajes más altos en las pruebas saber 11. Con lo anterior los beneficiarios de dichos créditos tendrían derecho a la financiación del 100% del valor de la matricula y a su vez un valor de sostenimiento que les permitía a los estudiantes tener una ayuda económica durante cada semestre.

Frente a las funciones y capacidad que tiene el Ministerio de Educación Nacional no se encuentra la administración de créditos o becas para la ciudadanía colombiana o

extranjera, de esta forma, acude al ICETEX, entidad reglamentada por medio de La Ley 1002 de 2005 que transforma al ICETEX en entidad financiera de naturaleza especial, la Ley 30 de 1992 y la Ley 1012 de 2006 faculta al ICETEX como administrador de recursos. Con lo anterior, las dos instituciones tomaron la decisión de crear un Fondo en Administración con el cual el Ministerio de Educación Nacional entrega un mandato al ICETEX en todo lo concerniente a la administración y operación de los recursos para el otorgamiento de créditos educativos condonables.

Como servidores públicos y conocedores de las estrategias públicas que realizan las entidades estatales frente al acceso, permanencia y graduación en la educación superior, acudimos a las dos instituciones en comento a través de un derecho de petición, para poner en conocimiento la intención de solicitar y usar la información necesaria del programa ser pilo paga, para realizar un análisis de los datos en calidad de estudiantes de la maestría Business Analytics; resultado de nuestra petición se obtuvo una base anonimizada de 39.998 beneficiarios del programa Ser Pilo Paga con sus condiciones socioeconómicas, académicas y demográficas en donde se informó el estado de los beneficiarios a la fecha: graduados, desertados, estudiando y que no utilizaron el beneficio.

Con dicha información se realiza un estudio de los estudiantes graduados y desertados, para determinar las variables que impactan o no, en la deserción académica. La condición de los beneficiarios a dicha fecha es:

**Tabla 2**

*Condición de los beneficiarios*

| **ESTADO** | **No.** |
|---|---|
| Condonados o graduados | 11.164 |
| Condonación muerte o invalidez | 61 |
| Cupo no utilizado | 2.041 |
| Estudiando | 6.749 |
| Inactivos desertores | 5.105 |
| Susceptibles condonaciones | 14.878 |
| **TOTAL** | **39.998** |

*Nota:* Muestra estado y numero de la población de ser pilo paga. Fuente: Elaboración Propia, (2022)

Dentro de las características de la información, nos entregan la data con fecha de corte al 30 de junio de 2022, haciendo la salvedad que estas características podrían variar a través del tiempo, por ello se aclara que el proyecto empresarial partirá de la base entregada, la cual es analizada y evaluada con respecto a la calidad de data.

## 6. Descripción de la situación estudio de caso y/o problemática empresarial y método y/o estrategia a aplicar para su solución

El programa ser pilo paga fue diseñado para ofrecer acceso, permanencia y graduación de la educación superior, a la población vulnerable con resultados sobresalientes en las pruebas saber 11; sin embargo, a pesar de este apoyo para acceder a la educación superior, se están presentado deserciones académicas de la población frente a las condiciones iniciales con las cuales se adjudicó el crédito educativo condonable.

La base anonimizada de los beneficiarios del programa ser pilo paga en sus diferentes versiones, cuenta con 27 variables para desarrollar el modelo, tal como se muestra a continuación.

**Tabla 3**

*Listado de Variables*

| No. | Variables |
|---|---|
| 1 | VERSIÓN |
| 2 | ESTRATO |
| 3 | EDAD |
| 4 | GENERO |
| 5 | DEPARTAMENTO_DE_FAMILIA |
| 6 | CIUDAD_DE_FAMILIA |
| 7 | EDUCACION_PADRE |
| 8 | EDUCACION_MADRE |
| 9 | OCUPACION_PADRE |
| 10 | OCUPACION_MADRE |
| 11 | INGRESO_FMILIAR_MENSUAL |
| 12 | DEPARTAMENTO_DE_COLEGIO |
| 13 | CIUDAD_DE_COLEGIO |
| 14 | COLEGIO |
| 15 | COLE_NATURALEZA |
| 16 | VALOR_PENSION_COLEGIO |
| 17 | COLE_JORNADA |
| 18 | PUNTAJE_SABER_11 |
| 19 | PUNTAJE_SISBENII |
| 20 | AREA_SISBENIII |
| 21 | SUBSIDIO_ASIGNADO |
| 22 | ORIGEN_DE_LA_IES |
| 23 | IES |
| 24 | DEPARTAMENTO_IES |
| 25 | CIUDAD_IES |
| 26 | PROGRAMA |
| 27 | ESTADO_DEL_BENEFICIARIO |

*Nota*, Cantidad de variables de la data. Fuente: Elaboración Propia, (2023)

Si bien es cierto que el Ministerio de Educación Nacional es la entidad del estado colombiano encargada de garantizar el acceso, permanencia y graduación de los colombianos en la educación superior, y que sus diversas estrategias lo han promovido así, no dejan de ser inciertos los factores por los que se presenta la deserción académica de los jóvenes en general, para el caso particular de la población vulnerable del programa ser pilo paga se evidencia un alto volumen de deserción, puesto que según las cifras entregadas por esta entidad al corte 30 de junio de 2022, los jóvenes que desertaron de este programa asciende a 5.105 de un total de 39.998 beneficiarios, lo cual equivale a un 12,7% del total de los beneficiarios.

Por esta razón, el modelo de predicción se construirá a partir de la población desertora que corresponde a 5.105 beneficiarios y a la población graduada que corresponde a 11.164 beneficiarios para así validar su efectividad. Adicionalmente, el modelo desarrollado se aplicará posteriormente a la población que se encuentra a fecha de 30 de junio de 2022 estudiando y que es susceptible de condonación, la cual asciende a 21.627 jóvenes, que representan el 54.07% del total de la población.

Sin embargo, antes de iniciar la construcción del modelo predictivo, se somete la data recibida a un proceso de validación y verificación de calidad, por lo que utilizamos las principales dimensiones para su evaluación, las cuales arrojan los siguientes resultados:

## Dimensión de Completitud

### Tabla 4

*Dimensión de Completitud*

| N. | VARIABLES | campos | registros totales | completitud |
|---|---|---|---|---|
| **1** | VERSIÓN | 39.998 | 39.998 | 100% |
| **2** | ESTRATO | 39.363 | 39.998 | 98% |
| **3** | EDAD | 39.363 | 39.998 | 98% |
| **4** | GENERO | 39.998 | 39.998 | 100% |
| **5** | DEPARTAMENTO_DE_FAMILIA | 39.363 | 39.998 | 98% |
| **6** | CIUDAD_DE_FAMILIA | 39.363 | 39.998 | 98% |
| **7** | EDUCACION_PADRE | 36.655 | 39.998 | 92% |
| **8** | EDUCACION_MADRE | 36.667 | 39.998 | 92% |
| **9** | OCUPACION_PADRE | 36.695 | 39.998 | 92% |
| **10** | OCUPACION_MADRE | 36.667 | 39.998 | 92% |
| **11** | INGRESO_FMILIAR_MENSUAL | 29.123 | 39.998 | 73% |
| **12** | DEPARTAMENTO_DE_COLEGIO | 39.998 | 39.998 | 100% |
| **13** | CIUDAD_DE_COLEGIO | 39.998 | 39.998 | 100% |
| **14** | COLEGIO | 39.992 | 39.998 | 100% |
| **15** | COLE_NATURALEZA | 36.730 | 39.998 | 92% |
| **16** | VALOR_PENSION_COLEGIO | 29.029 | 39.998 | 73% |
| **17** | COLE_JORNADA | 36.730 | 39.998 | 92% |
| **18** | PUNTAJE_SABER_11 | 39.996 | 39.998 | 100% |
| **19** | PUNTAJE_SISBENII | 39.716 | 39.998 | 99% |
| **20** | AREA_SISBENIII | 39.716 | 39.998 | 99% |
| **21** | SUBSIDIO_ASIGNADO | 39.363 | 39.998 | 98% |
| **22** | ORIGEN_DE_LA_IES | 39.363 | 39.998 | 98% |
| **23** | IES | 39.363 | 39.998 | 98% |
| **24** | DEPARTAMENTO_IES | 39.363 | 39.998 | 98% |
| **25** | CIUDAD_IES | 39.363 | 39.998 | 98% |
| **26** | PROGRAMA | 39.363 | 39.998 | 98% |
| **27** | ESTADO_DEL_BENEFICIARIO | 39.998 | 39.998 | 100% |

*Nota.* Información de datos completos de la data. Fuente: Elaboración Propia, (2023)

## Dimensión validez

### Tabla 5

*Dimensión Validez*

| N. | VARIABLES | FORMATO VALIDO | REQUERIMIENTOS VALIDEZ | REGISTROS TOTALES | VALIDEZ |
|---|---|---|---|---|---|
| **1** | VERSIÓN | REGISTRO ALFANUMERICO | 39.998 | 39.998 | 100% |
| **2** | ESTRATO | REGISTRO NUMERICO DEL 1 AL 6 | 39.363 | 39.998 | 98% |
| **3** | EDAD | REGISTRO NUMERICO DEL 15 AL 70 | 39.363 | 39.998 | 98% |
| **4** | GENERO | REGISTRO ALFABETICO (M O F) | 39.998 | 39.998 | 100% |
| **5** | DEPARTAMENTO_DE_FAMILIA | REGISTRO CATEGORICO DE LOS 32 DEPARTAMENTOS DE COLOMBIA Y DISTRITO CAPITAL | 39.363 | 39.998 | 98% |
| **6** | CIUDAD_DE_FAMILIA | REGISTRO CATEGORICO Y ALFABETICO DE LOS MUNICIPIOS DE COLOMBIA | 39.363 | 39.998 | 98% |
| **7** | EDUCACION_PADRE | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 36.655 | 39.998 | 92% |
| **8** | EDUCACION_MADRE | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 36.667 | 39.998 | 92% |

| N. | VARIABLES | FORMATO VALIDO | REQUERIMIENTOS VALIDEZ | REGISTROS TOTALES | VALIDEZ |
|---|---|---|---|---|---|
| **9** | OCUPACION_PADRE | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 36.695 | 39.998 | 92% |
| **10** | OCUPACION_MADRE | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 36.667 | 39.998 | 92% |
| **11** | INGRESO_FMILIAR_MENSUAL | REGISTRO ALFABETICO Y/O NUMERICO SIN LIMITE DE CARACTERES | 29.123 | 39.998 | 73% |
| **12** | DEPARTAMENTO_DE_COLEGIO | REGISTRO CATEGORICO DE LOS 32 DEPARTAMENTOS DE COLOMBIA Y DISTRITO CAPITAL | 39.998 | 39.998 | 100% |
| **13** | CIUDAD_DE_COLEGIO | REGISTRO CATEGORICO Y ALFABETICO DE LOS MUNICIPIOS DE COLOMBIA | 39.998 | 39.998 | 100% |
| **14** | COLEGIO | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 39.992 | 39.998 | 100% |
| **15** | COLE_NATURALEZA | REGISTRO CATEGORICO (OFICIAL, NO OFICIAL) | 36.730 | 39.998 | 92% |
| **16** | VALOR_PENSION_COLEGIO | REGISTRO ALFABETICO Y/O NUMERICO SIN LIMITE DE CARACTERES | 29.029 | 39.998 | 73% |
| **17** | COLE_JORNADA | REGISTRO CATEGORICO (JORNADAS COLEGIOS) | 36.730 | 39.998 | 92% |
| **18** | PUNTAJE SABER 11 | REGISTRO NUMERICO ENTRE 300 Y 500 | 39.996 | 39.998 | 100% |
| **19** | PUNTAJE_SISBENII | REGISTRO NUMERICO ENTRE 0 Y 100 | 39.716 | 39.998 | 99% |
| **20** | AREA_SISBENIII | REGISTRO NUMERICO ENTRE 1 Y 3 | 39.716 | 39.998 | 99% |
| **21** | SUBSIDIO ASIGNADO | REGISTRO CATEGORICO (1, 1.5, 2, 2.5, 4 Y 4.5) | 39.363 | 39.998 | 98% |
| **22** | ORIGEN DE LA IES | REGISTRO CATEGORICO (CARÁCTER ESPECIAL, OFICIAL Y PRIVADO) | 39.363 | 39.998 | 98% |
| **23** | IES | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 39.363 | 39.998 | 98% |
| **24** | DEPARTAMENTO IES | REGISTRO CATEGORICO DE LOS 32 DEPARTAMENTOS DE COLOMBIA Y DISTRITO CAPITAL | 39.363 | 39.998 | 98% |
| **25** | CIUDAD IES | REGISTRO CATEGORICO Y ALFABETICO DE LOS MUNICIPIOS DE COLOMBIA | 39.363 | 39.998 | 98% |
| **26** | PROGRAMA | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 39.363 | 39.998 | 98% |
| **27** | ESTADO_DEL_BENEFICIARIO | REGISTRO ALFABETICO SIN LIMITE DE CARACTERES | 39.998 | 39.998 | 100% |
| **VALIDEZ BASE DE BENEFICIARIOS** | | | **34.208** | **39.998** | **86%** |

*Nota.* Explica grado de conformidad con formato, tipado y rango. Fuente: Elaboración Propia, (2023)

## Dimensión puntualidad

Esta dimensión no se puede calcular, por cuanto se solicitó la información al corte de 30 de junio de 2022, y fue suministrada el 08 de agosto de 2022, sin embargo, es de señalar que las variables aquí tratadas no tienen relación con tiempos de emisión.

**Dimensión Unicidad**

De las 27 variables evaluadas se determina que no existe unicidad, ya que pueden ser categóricas y pueden presentar la misma condición varios de los beneficiarios. Sin embargo, se realizó la validación de registros duplicados en la totalidad de la base y se constató que son registros únicos.

**Dimensión Precisión**

**Tabla 6**

*Dimensión precisión versión ser pilo paga*

| VERSIÓN | REGISTROS PRECISOS | REGISTROS TOTALES | PRECISION |
|---|---|---|---|
| **SPP1** | 10.142 | 10.142 | 100% |
| **SPP2** | 12.751 | 12.751 | 100% |
| **SPP3** | 9.086 | 9.086 | 100% |
| **SPP4** | 8.019 | 8.019 | 100% |

*Nota,* Grado en que los datos representan la realidad en la versión. Fuente: Elaboración Propia, (2023)

**Tabla 7**

*Dimensión precisión por área de Sisbén*

| **DIMENSION PRECISION** | | | |
|---|---|---|---|
| **AREA SISBEN** | REGISTROS PRECISOS | REGISTROS TOTALES | PRECISION |
| **1** | 16.784 | 16.784 | 100% |
| **2** | 19.620 | 19.620 | 100% |
| **3** | 3.312 | 3.312 | 100% |

*Nota,* Grado en que los datos representan la realidad en área Sisbén. Fuente: Elaboración Propia, (2023)

**Tabla 8**

*Condiciones de las versiones de SPP Vs. Prueba Saber11 y Sisbén*

| VERSIÓN | SABER 11 | PUNTAJE MAXIMO SISBEN |
|---|---|---|
| **SPP1** | Mayor o igual 310 | AREA 1: 57,21<br>AREA 2: 56,32<br>AREA 3: 40,75 |
| **SPP2** | Mayor o igual 318 | |
| **SPP3** | Mayor o igual 342 | |
| **SPP4** | Mayor o igual 348 | |

*Nota,* Puntajes exigidos por versión. Fuente: Elaboración Propia, (2023)

**Dimensión consistencia**

**Tabla 9**

*Dimensión de consistencia*

| N. | VARIABLES | REGISTROS CONSISTENTES | REGISTROS TOTALES | CONSISTENCIA |
|---|---|---|---|---|
| 1 | DEPARTAMENTO_DE_FAMILIA | 33.640 | 39.998 | 84% |
| 2 | CIUDAD_DE_FAMILIA | 26.003 | 39.998 | 65% |
| 3 | DEPARTAMENTO_DE_COLEGIO | 34.079 | 39.998 | 85% |
| 4 | CIUDAD_DE_COLEGIO | 26.643 | 39.998 | 67% |
| 5 | DEPARTAMENTO IES | 26.061 | 39.998 | 65% |
| 6 | CIUDAD IES | 24.093 | 39.998 | 60% |

*Nota,* Nivel en que dispone información para consulta. Fuente: Elaboración Propia, (2023)

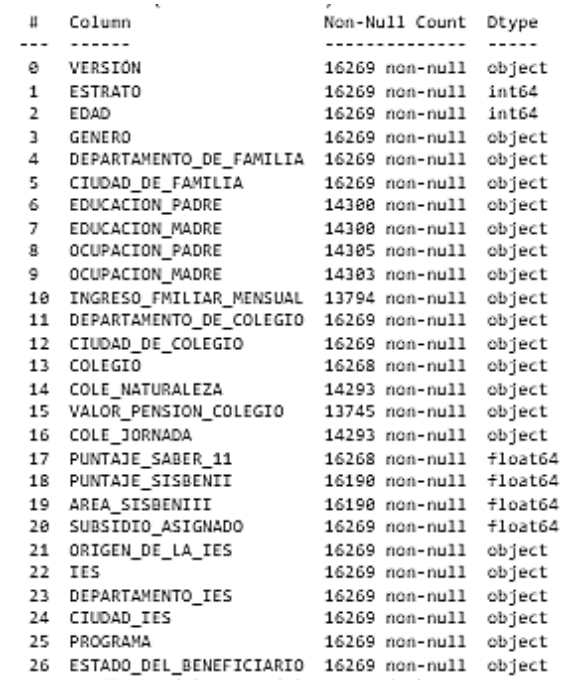
En cuanto a la dimensión de consistencia encontramos que frente a los resultados obtenidos las diferencias obedecen a las tildes (acentos) no registrados en la base de datos, vs base DIVIPOLA, y adicional a ello por la codificación diferente en la base de datos del departamento Distrito capital. Se tomó la base llamada DIVIPOLA, que contempla los nombres de los departamentos y municipios de Colombia según codificación del DANE, versus la información de la base.

Ahondado en la data y específicamente en los beneficiarios que son objeto de análisis para la creación del modelo de predicción, recurrimos a la analítica descriptiva a través de colaboratory de Google, en el cual se obtuvo los siguientes resultados:

## 6.1.Identificación de variables y revisión de datos nulos:

**Figura 3**

*Listado de variables con datos nulos*



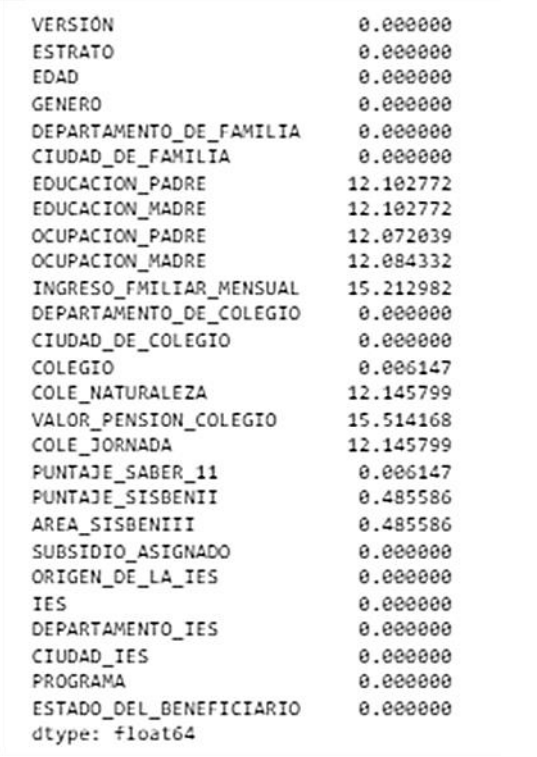
| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | VERSIÓN | 16269 non-null | object |
| 1 | ESTRATO | 16269 non-null | int64 |
| 2 | EDAD | 16269 non-null | int64 |
| 3 | GENERO | 16269 non-null | object |
| 4 | DEPARTAMENTO_DE_FAMILIA | 16269 non-null | object |
| 5 | CIUDAD_DE_FAMILIA | 16269 non-null | object |
| 6 | EDUCACION_PADRE | 14300 non-null | object |
| 7 | EDUCACION_MADRE | 14300 non-null | object |
| 8 | OCUPACION_PADRE | 14305 non-null | object |
| 9 | OCUPACION_MADRE | 14303 non-null | object |
| 10 | INGRESO_FMILIAR_MENSUAL | 13794 non-null | object |
| 11 | DEPARTAMENTO_DE_COLEGIO | 16269 non-null | object |
| 12 | CIUDAD_DE_COLEGIO | 16269 non-null | object |
| 13 | COLEGIO | 16268 non-null | object |
| 14 | COLE_NATURALEZA | 14293 non-null | object |
| 15 | VALOR_PENSION_COLEGIO | 13745 non-null | object |
| 16 | COLE_JORNADA | 14293 non-null | object |
| 17 | PUNTAJE_SABER_11 | 16268 non-null | float64 |
| 18 | PUNTAJE_SISBENII | 16190 non-null | float64 |
| 19 | AREA_SISBENIII | 16190 non-null | float64 |
| 20 | SUBSIDIO_ASIGNADO | 16269 non-null | float64 |
| 21 | ORIGEN_DE_LA_IES | 16269 non-null | object |
| 22 | IES | 16269 non-null | object |
| 23 | DEPARTAMENTO_IES | 16269 non-null | object |
| 24 | CIUDAD_IES | 16269 non-null | object |
| 25 | PROGRAMA | 16269 non-null | object |
| 26 | ESTADO_DEL_BENEFICIARIO | 16269 non-null | object |

Fuente: Elaboración propia, (2023)

En la figura anterior se muestra las variables que registran datos nulos. Frente a la revisión realizada, encontramos que se presentan en algunos datos nulos, siendo las más alta, la variable No. 15 con 2.524 nulos, lo cual es un punto por considerar en el momento de la experimentación y definición del modelo, dado que la regresión logística requiere la completitud de los datos.

**Figura 4**

*Porcentaje de datos nulos en las variables*



```
VERSIÓN                    0.000000
ESTRATO                    0.000000
EDAD                       0.000000
GENERO                     0.000000
DEPARTAMENTO_DE_FAMILIA    0.000000
CIUDAD_DE_FAMILIA          0.000000
EDUCACION_PADRE           12.102772
EDUCACION_MADRE           12.102772
OCUPACION_PADRE           12.072039
OCUPACION_MADRE           12.084332
INGRESO_FMILIAR_MENSUAL   15.212982
DEPARTAMENTO_DE_COLEGIO    0.000000
CIUDAD_DE_COLEGIO          0.000000
COLEGIO                    0.006147
COLE_NATURALEZA           12.145799
VALOR_PENSION_COLEGIO     15.514168
COLE_JORNADA              12.145799
PUNTAJE_SABER_11           0.006147
PUNTAJE_SISBENII           0.485586
AREA_SISBENIII             0.485586
SUBSIDIO_ASIGNADO          0.000000
ORIGEN_DE_LA_IES           0.000000
IES                        0.000000
DEPARTAMENTO_IES           0.000000
CIUDAD_IES                 0.000000
PROGRAMA                   0.000000
ESTADO_DEL_BENEFICIARIO    0.000000
dtype: float64
```

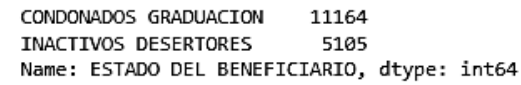
*Nota,* Variables con porcentajes de datos nulos. Fuente: Elaboración propia, (2023)

En la figura anterior, se describe el porcentaje de datos nulos que existen en cada una de las variables, teniendo un porcentaje de nulos de las variables en un rango de 12% a 15% siendo representativo para la experimentación que se quiere realizar en búsqueda del modelo más efectivo, toda vez que influirán si se desea utilizar alguna de estas variables en la perdida de información en la regresión logística.

**Cantidad de graduados y desertores**:

**Figura 5**

***# Graduados y # desertores.***



```
CONDONADOS GRADUACION    11164
INACTIVOS DESERTORES      5105
Name: ESTADO DEL BENEFICIARIO, dtype: int64
```

*Nota,* Conteo de beneficiarios por estado. Fuente Elaboración propia, (2023)

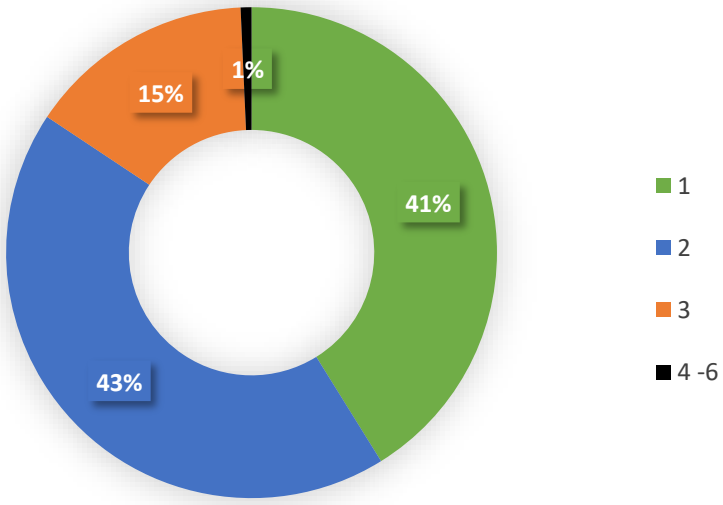
De los 16.269 beneficiarios, 11.164 se encuentran graduados o condonados; por otra parte, se cuenta con 5.105 estudiantes desertores o inactivos en el programa Ser Pilo Paga.

## 6.2.Visualizaciones de la data

### Estrato

La mayoría de los beneficiarios se encuentran en los estratos 1 y 2, con una participación del 84%, es decir, 13.722 beneficiarios. Mientras que, para el estrato 3 se tiene una participación del 15% con 2.433 beneficiarios. Los beneficiarios de los estratos 4,5 y 6 tienen una participación menor al 1% que corresponden a 116 casos atípicos de ciudades con estratificación diferente.

**Figura 6**
*Distribución por estrato*



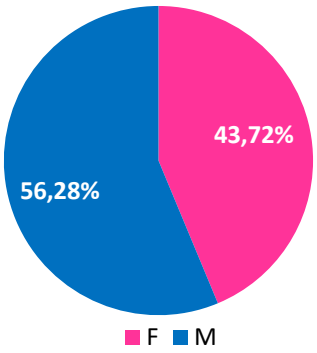

*Nota*, Beneficiarios por estrato. Fuente: Elaboración propia (2023)

**Género**

La población con mayor participación es el género masculino con un total de 9.156 beneficiarios que corresponden a un 56,28%; por otra parte, el género femenino presenta un total de 7.113 que corresponde a un 43,72%

**Figura 7**

*Distribución por género*

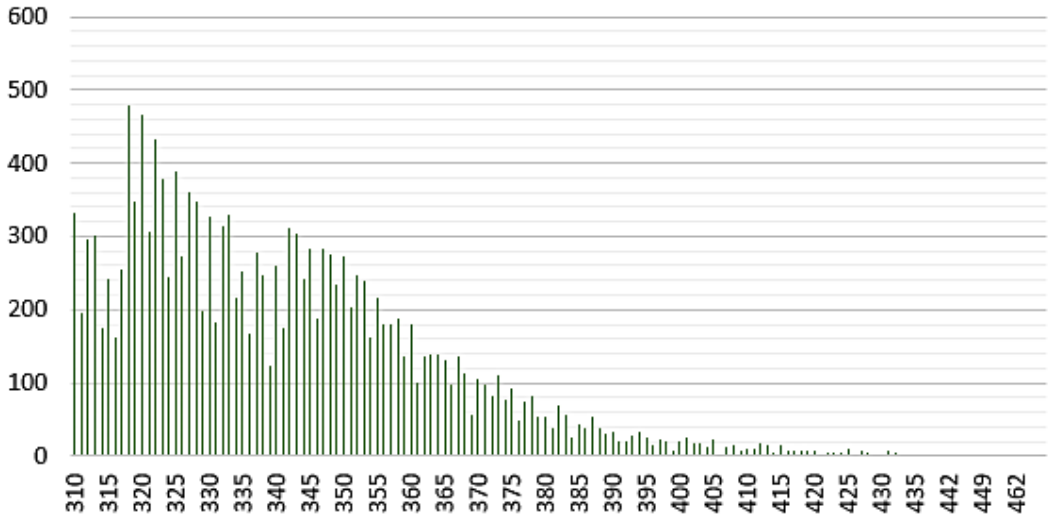



*Nota,* Número de beneficiarios por género. Fuente: Elaboración propia, (2023)

**Puntaje Saber 11**

**Figura 8**

*Histograma puntaje saber 11*



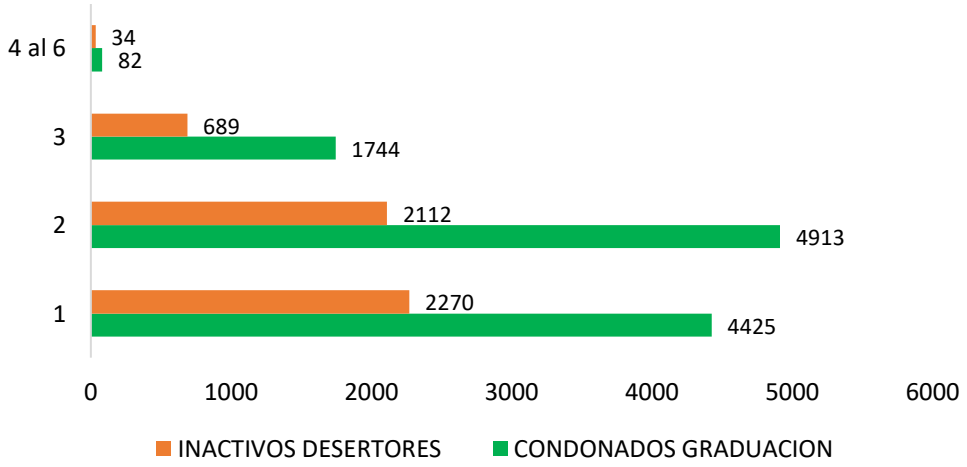

*Nota,* Número de Beneficiarios por puntaje Saber11. Fuente: Elaboración Propia, (2023)

Se evidencia que los puntajes de pruebas saber 11 inician en 310 puntos, presentando su mayor concentración desde este puntaje y hasta los 370 puntos aproximadamente, a partir de este puntaje se encuentran los beneficiarios con los mejores puntajes obtenidos. Así mismo, se presentan algunos datos atípicos cercanos a 470 puntos.

**Graduación o Deserción por estrato**

**Figura 9**

*Gráfico de barras condonados graduación vs inactivos desertores por estrato*



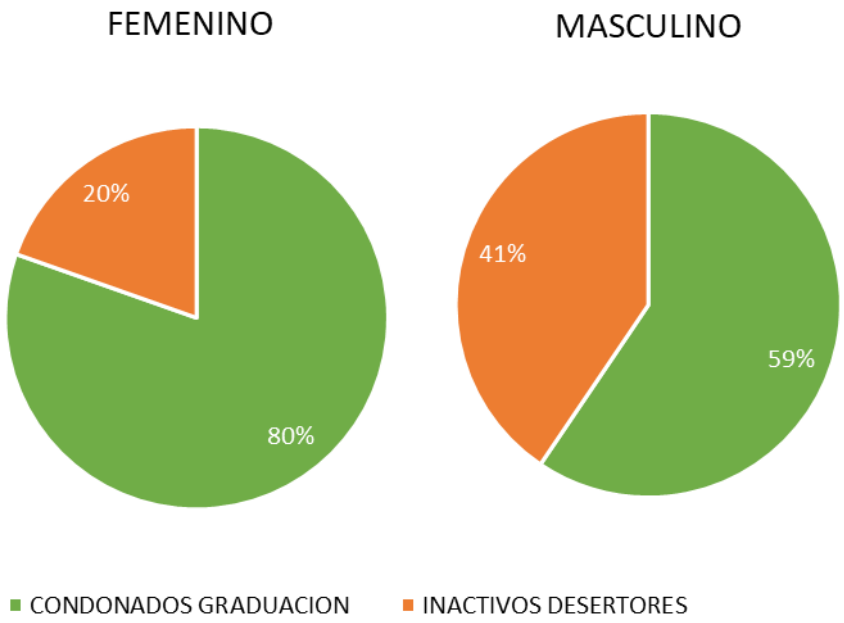

*Nota*, Condonados vs Desertores por estrato. Fuente: Elaboración Propia, (2023)

Frente a la comparación de estado del beneficiario (Graduado o desertado) con respecto al estrato, podemos evidenciar un comportamiento similar entre estrato 1 y 2 con respecto a las dimensiones de cada una de las dos situaciones, sin embargo, para el caso de los graduados se han presentado más en el estrato 2 y por el contrario la mayor deserción se presenta en el estrato 1.

**Graduación o Deserción por género**

**Figura 10**

*Condonados graduación vs inactivos desertores por género*




*Nota*, Condonados Vs desertores por género. Fuente: Elaboración propia, (2023)

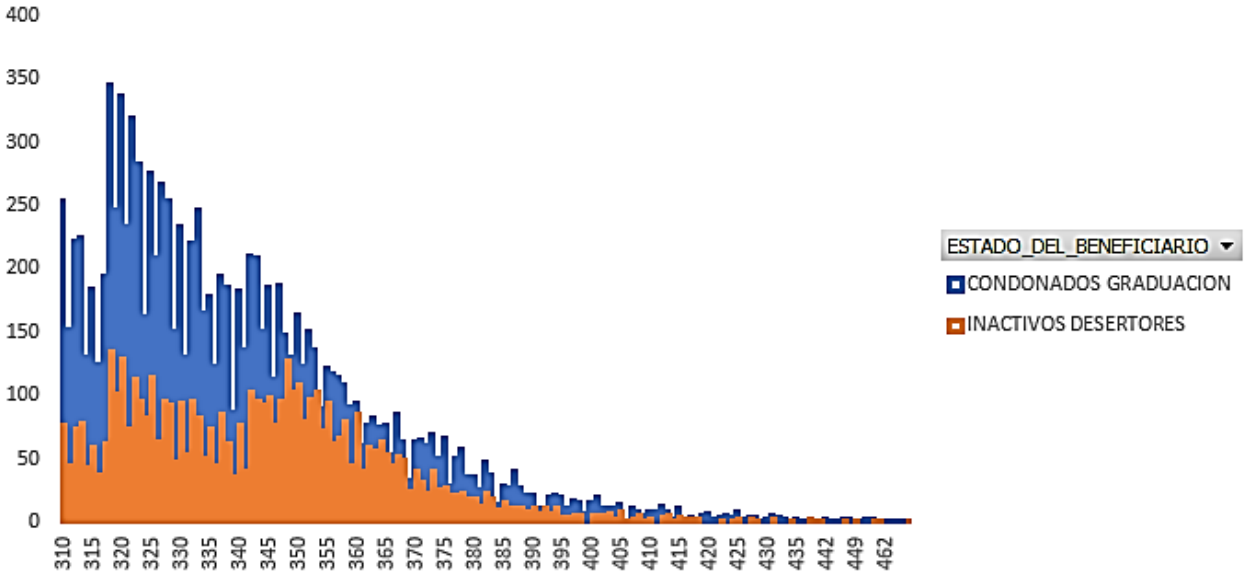
El comportamiento de las mujeres que logran una condonación y graduación es superior a la de los hombres en un 21%. En cuanto a la deserción o inactivos, los hombres representan un mayor porcentaje de participación con relación a las mujeres.

**Graduación o Deserción por puntaje saber 11**

Frente a los beneficiarios inactivos o desertores se muestra la mayor deserción en los puntajes inferiores a 360 puntos, mientras que los mayores puntajes de la población condonada o graduada se encuentra en el rango de 320 a 345 puntos aproximadamente.

**Figura 11**

*Histograma Graduación o Deserción por Puntaje saber 11*

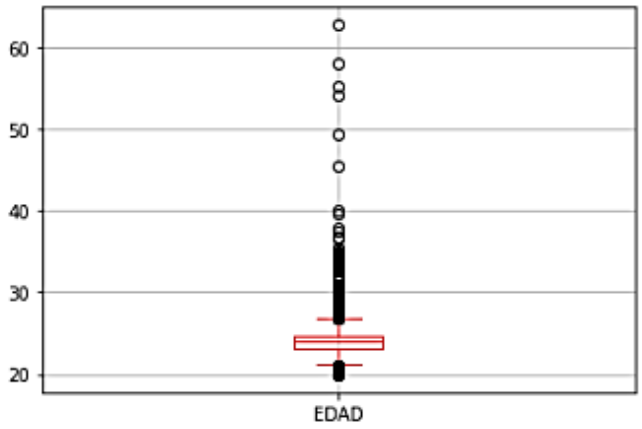



*Nota,* Graduados Vs Desertados por Puntaje Saber11. Fuente: Elaboración propia, (2023)

**Distribución por edad**

**Figura 12**

*Caja de Bigotes por edad*




*Nota,* Outliers y concentración por edad. Fuente: Elaboración propia, (2023)

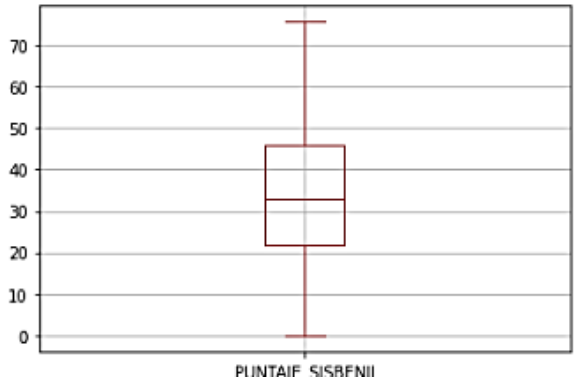
Con respecto a la edad, tenemos unos beneficiarios o outliers que tienen una edad menor a 21 años y superior a 25 años aproximadamente.

**Distribución por Puntaje Sisbén II**

Para el caso puntaje de Sisbén II, encontramos que el 75% de la población tiene un puntaje inferior a 45 puntos aproximadamente o no cuentan con puntaje.

**Figura 13**

*Caja de Bigotes por puntaje Sisbén*

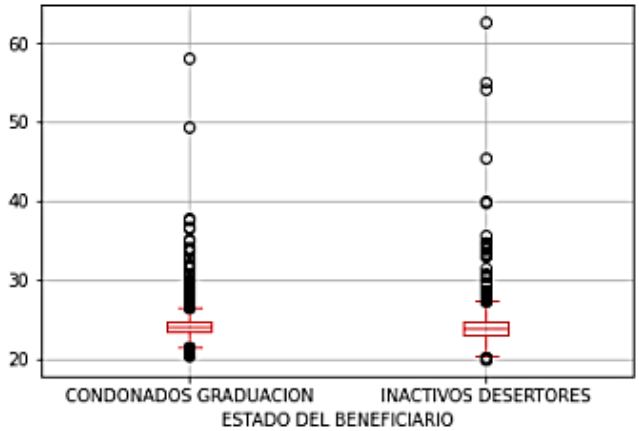



*Nota*, Concentración de la población según Sisbén. Fuente: Elaboración Propia, (2023).

**Distribución de graduados y desertados por edad.**

**Figura 14**

*Caja de Bigotes graduados y desertados por edad.*




*Nota*, Concentración y outliers condonados vs graduados. Fuente: Elaboración propia, (2023).

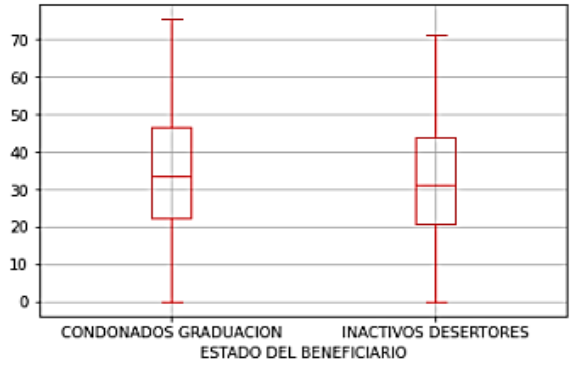
Separando la situación de edad por el estado del beneficiario (Graduado o desertado), el rango de edad de los desertados es un poco más amplio, mas no representa

una gran diferencia con los graduados, para las dos situaciones se siguen presentando outliers tanto por el extremo inferior y superior.

**Distribución de graduados y desertados por Puntaje Sisbén** II.

**Figura 15**

Caja de Bigotes graduados y desertados por Puntaje Sisbén II.




*Nota*, Concentración graduados vs desertores Puntaje SisbénII, Fuente: Elaboración propia, (2023)

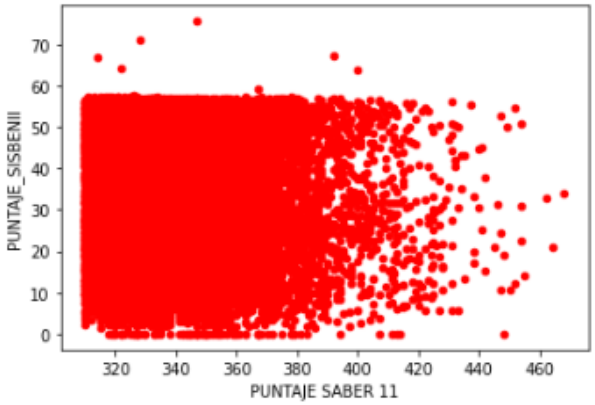
Separando la situación de Puntaje Sisbén II por la situación del beneficiario (Graduado o desertado), se evidencia un comportamiento similar con pequeñas diferencias como que la mediana de los graduados es un poco mayor a los desertados, así como su primer y tercer cuartil.

**Dispersión puntaje saber 11 vs puntaje Sisbén II**

Encontramos en la dispersión, que los beneficiarios presentan una concentración en un puntaje igual o inferior a 410 puntos de pruebas saber 11 y un puntaje de Sisbén II igual o inferior a 58 puntos aproximadamente.

**Figura 16**

*Gráfico Dispersión puntaje saber 11 vs puntaje Sisbén II*

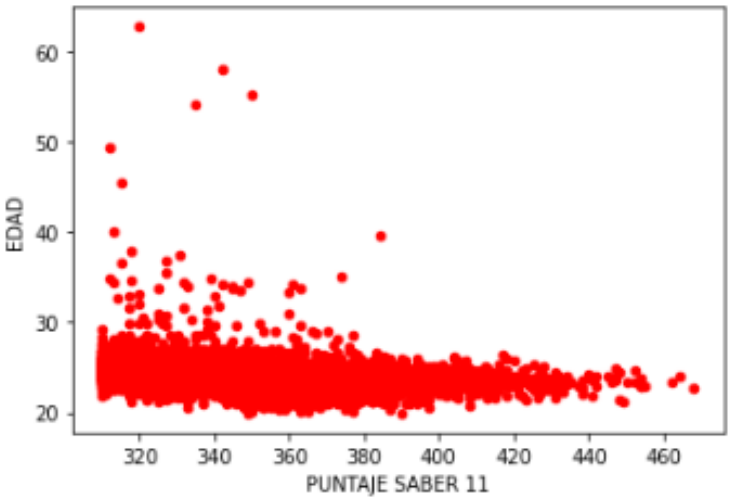



*Nota,* Relación entre puntaje saber11 y puntaje sisbénII. Fuente: Elaboración propia, (2023)

**Dispersión Edad vs puntaje saber 11**

**Figura 17**

*Gráfico dispersión Edad vs puntaje saber 11*




*Nota,* Relación entre edad y puntaje Saber11. Fuente: Elaboración propia, (2023)

No se evidencia una relación directa entre la edad promedio con los puntajes de las pruebas saber 11, toda vez que tenemos un rango muy amplio de puntajes que se puede presentar. Sin embargo, encontramos que las personas mayores a 30 años no tienen los mejores

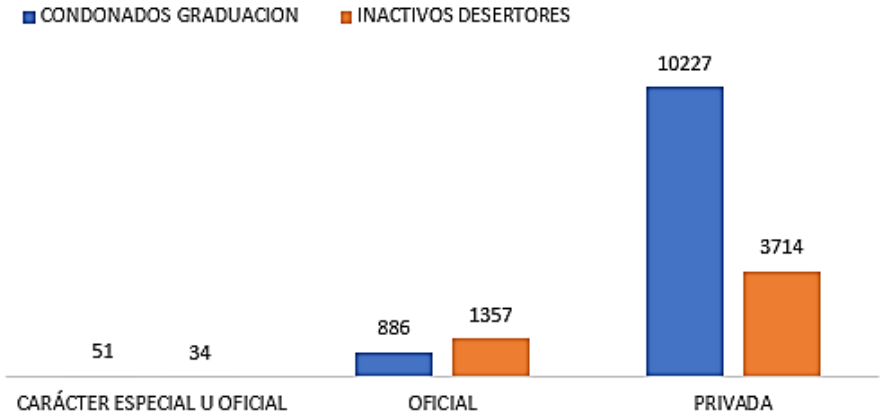
puntajes, excepto un caso que tiene aproximadamente 385 puntos, siendo el máximo en este rango de edad.

**Graduación o deserción por Carácter de la IES.**

Analizando la situación del comportamiento de graduados y desertados por Instituciones de Educación Superior, encontramos que para las IES oficiales (públicas) se presentan mayores deserciones que graduaciones, lo que no pasa con las IES privadas y de carácter especial (Instituciones de formación militar).

**Figura 18**

*Gráfico de barras graduación o deserción por carácter IES.*




Fuente: Elaboración propia, (2023)

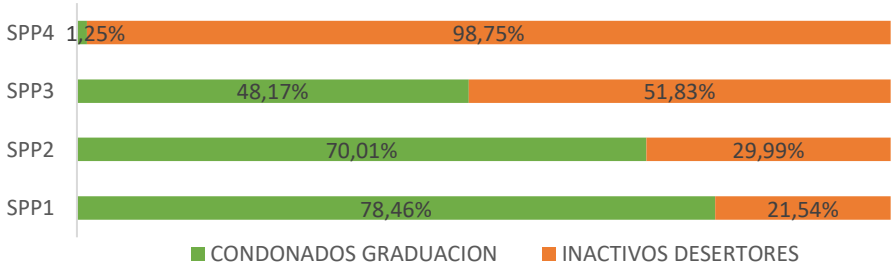
Se puede evidenciar que los estudiantes condonados o graduados pertenecientes a Instituciones de Educación superior Privada, representan mayor proporción frente a los inactivos desertores. Mientras que en las instituciones de educación superior oficiales los inactivos o desertores son mayores a los condonados o graduados.

**Graduación o deserción por Versión del programa Ser Pilo Paga.**

Para las versiones 1 y 2 del programa ser pilo paga se encuentra un comportamiento similar, en el que los beneficiarios graduados superan los desertados, para la versión 3 el número de estudiantes graduados y desertados son similares representando cada uno casi un 50%. Por último, la versión 4 presentan un comportamiento particular dado que los desertores corresponden al 98.75% mientras que los graduados solo representan un 1.25%.

**Figura 19**

*Gráfico de Barras graduación o deserción por Versión SPP.*

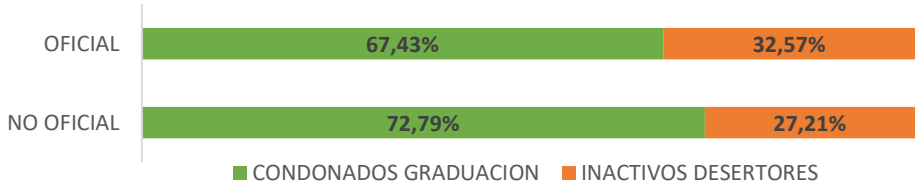



*Nota,* Grupos de condonados vs desertados por versión SPP. Fuente: Elaboración: Fuente propia, (2023)

En la gráfica anterior, se muestra que la versión SPP4 tuvo la mayor proporción de desertados. En cuanto al mayor numero de condonados se presentó en la versión SPP1.

**Figura 20**

*Gráfico de barras naturaleza del colegio Vs estado del beneficiario*

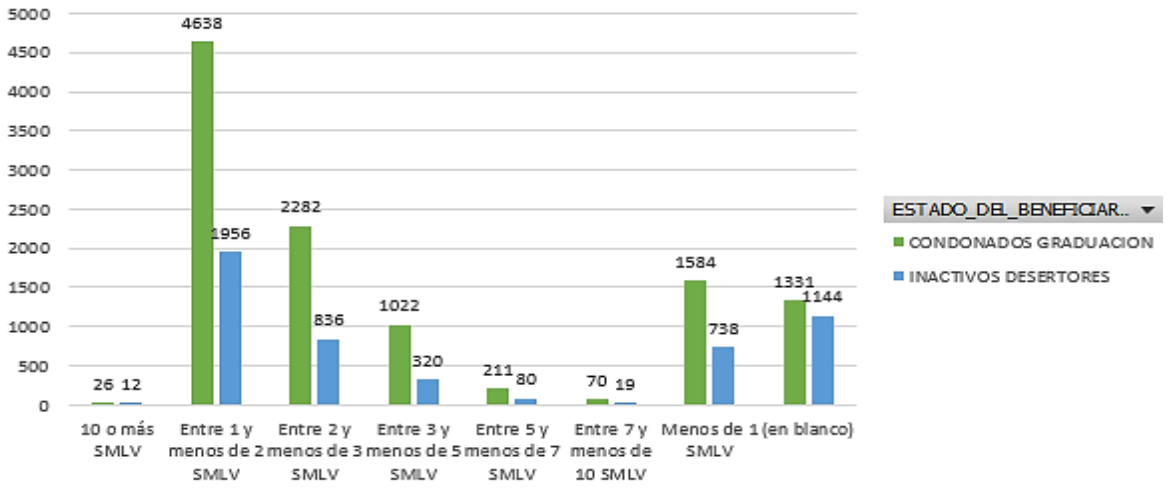



*Nota*, Comparativo variable naturaleza del colegio entre condonados y desertados. Fuente: Elaboración propia (2023).

En la figura anterior, se evidencia que los beneficiarios que son egresados de colegios no oficiales presentan un mayor porcentaje de graduación frente a los de colegio oficial representado en un 5.36%.

**Figura 21**

*Ingreso familiar VS Estado del beneficiario*




*Nota*, Comparativo ingreso familiar condonados vs desertores. Fuente: Elaboración propia, (2023)

En la figura anterior, se evidencia que los estudiantes que están en su mayoría graduados, su familia tenía un ingreso entre 1 y 2 SMMLV, por otra parte, para los

desertados se presentó en gran medida, que tenían un ingreso mínimo que para este caso es menor a 1 SMMLV.

**Figura 22**

*Tabla de Contingencia área Sisbén II Vs estado del Beneficiario*

| ESTADO_DEL_BENEFICIARIO | CONDONADOS | GRADUACION | INACTIVOS | DESERTORES |
|---|---|---|---|---|
| AREA_SISBENIII | | | | |
| 1.0 | | 0.710426 | | 0.289574 |
| 2.0 | | 0.668778 | | 0.331222 |
| 3.0 | | 0.661699 | | 0.338301 |

*Nota,* relación entre área Sisbén II y estado del beneficiario. Fuente: Elaboración propia, (2023).

En la figura anterior, no se encuentra una proporción relevante frente al área Sisbén II al que pertenecen los estudiantes, el porcentaje de graduación está en un rango entre 71,04% y 66,16%, siendo el más alto para los estudiantes del área de Sisbén número 1.; Para el caso de los desertados el rango es entre 28,95% y 33,83% siendo el más alto para el área de Sisbén número 3.

**Figura 23**

*Tabla de Contingencia Colegio Naturaleza Vs. Estado del Beneficiario*

| ESTADO_DEL_BENEFICIARIO | CONDONADOS | GRADUACION | INACTIVOS | DESERTORES |
|---|---|---|---|---|
| COLE_NATURALEZA | | | | |
| NO OFICIAL | | 0.727878 | | 0.272122 |
| OFICIAL | | 0.674308 | | 0.325692 |

*Nota,* relación entre variable colegio Naturaleza vs estado del beneficiario. Fuente: Elaboración propia, (2023)

Desde el punto proporcional de los datos, se evidencia que de los estudiantes que son egresados de colegios no oficiales existe un mayor porcentaje de graduación con respecto a colegios oficiales, lo que conllevo decir que proporcionalmente se presenta más deserciones en los estudiantes graduados de colegios oficiales.

**Contingencia Porcentual graduados y desertores por departamento de origen.**

Frente al comportamiento de graduados y desertados por departamento, encontramos que los territorios que más presentan graduados son Atlántico (79,6%), Bolívar (78,2), Vaupés y Vichada (75%), estos dos últimos dado los pocos beneficiarios que presentan. Ahora los departamentos que presentan mayor deserción son Amazonas (57%,1), Putumayo (55,1%), Guainía y Guaviare (50%), estos dos últimos dado los pocos beneficiarios que presentan.

**Figura 24**

*Tabla contingencia por departamento origen.*

| ESTADO DEL BENEFICIARIO | CONDONADOS | GRADUACION | INACTIVOS | DESERTORES |
|---|---|---|---|---|
| DEPARTAMENTO_DE_FAMILIA | | | | |
| AMAZONAS | | 0.428571 | | 0.571429 |
| ANTIOQUIA | | 0.665370 | | 0.334630 |
| ARAUCA | | 0.622951 | | 0.377049 |
| ATLANTICO | | 0.796109 | | 0.203891 |
| BOLIVAR | | 0.782012 | | 0.217988 |
| BOYACA | | 0.680357 | | 0.319643 |
| CALDAS | | 0.670051 | | 0.329949 |
| CAQUETA | | 0.560976 | | 0.439024 |
| CASANARE | | 0.573276 | | 0.426724 |
| CAUCA | | 0.545894 | | 0.454106 |
| CESAR | | 0.736655 | | 0.263345 |
| CHOCO | | 0.666667 | | 0.333333 |
| CORDOBA | | 0.625205 | | 0.374795 |
| CUNDINAMARCA | | 0.725205 | | 0.274795 |
| DISTRITO CAPITAL | | 0.737711 | | 0.262289 |
| GUAINIA | | 0.500000 | | 0.500000 |
| GUAVIARE | | 0.500000 | | 0.500000 |
| HUILA | | 0.634518 | | 0.365482 |
| LA GUAJIRA | | 0.672414 | | 0.327586 |
| MAGDALENA | | 0.722222 | | 0.277778 |
| META | | 0.645390 | | 0.354610 |
| NARIÑO | | 0.545455 | | 0.454545 |
| NORTE DE SANTANDER | | 0.662016 | | 0.337984 |
| PUTUMAYO | | 0.448819 | | 0.551181 |
| QUINDIO | | 0.598039 | | 0.401961 |
| RISARALDA | | 0.653061 | | 0.346939 |
| SAN ANDRES | | 0.727273 | | 0.272727 |
| SANTANDER | | 0.703236 | | 0.296764 |
| SUCRE | | 0.673740 | | 0.326260 |
| TOLIMA | | 0.654639 | | 0.345361 |
| VALLE DEL CAUCA | | 0.655681 | | 0.344319 |
| VAUPES | | 0.750000 | | 0.250000 |
| VICHADA | | 0.750000 | | 0.250000 |

*Nota,* Relación en porcentajes condonados vs desertados. Fuente: Elaboración propia, (2023)

**Contingencia Porcentual graduados y desertores por origen de la IES.**

**Figura 25**

*Tabla contingencia por origen de la Institución de Educación Superior.*

| ESTADO DEL BENEFICIARIO | CONDONADOS | GRADUACION | INACTIVOS | DESERTORES |
|---|---|---|---|---|
| ORIGEN DE LA IES | | | | |
| CARÁCTER ESPECIAL U OFICIAL | | 0.600000 | | 0.400000 |
| OFICIAL | | 0.395007 | | 0.604993 |
| PRIVADA | | 0.733592 | | 0.266408 |

*Nota,* Relación en porcentajes condonados vs desertores por IES. Fuente: Elaboración propia, (2023)

Complementado el ejercicio de validación por tipo de IES, encontramos que el 73,3% de los estudiantes de IES privadas se gradúan y por el contrario el 60,4% de los estudiantes de IES publicas presentan deserción.

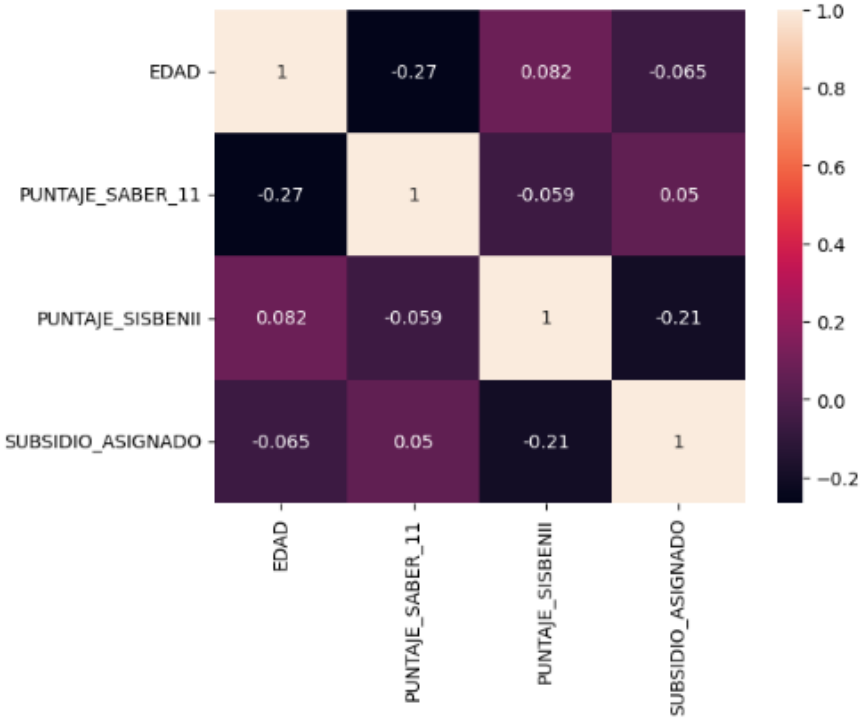
**Correlación de Variables**

Se realiza un primer acercamiento a la identificación de cuales variables son más relevantes frente al estado del beneficiario y como esas variables se correlación con las otras.

Dado lo anterior, se realizó una matriz de correlación de las variables independientes obteniendo el siguiente resultado:

**Figura 26**

*Matriz de correlación variables Independientes*




*Nota,* Relación variables. Fuente: Elaboración propia, *(2023)*

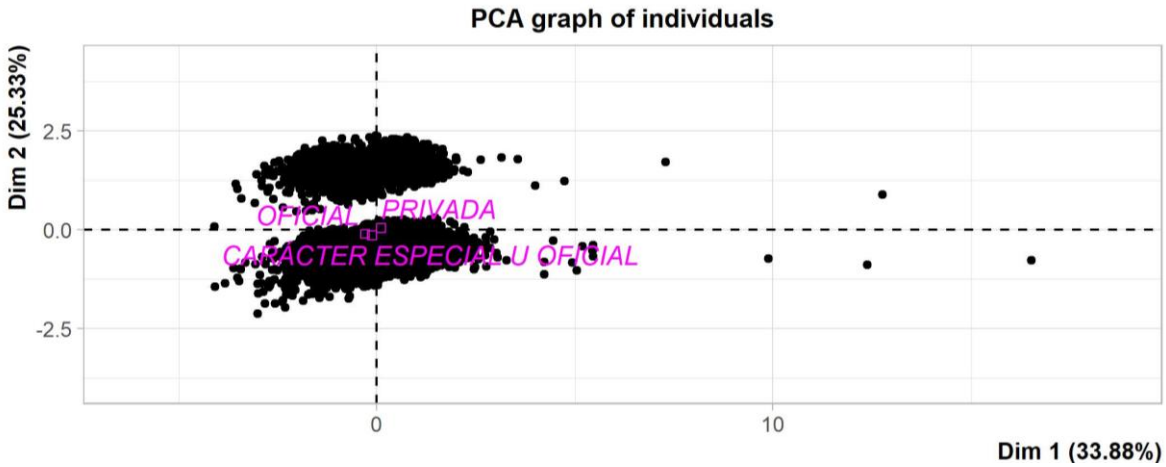
Al analizar la matriz se evidencia que las mayores correlaciones se encuentran, entre las variables independientes entre Puntaje Saber 11 y Edad, siendo esta de manera inversa y no tan relevante. En donde a su vez se muestra una poca correlación entre las variables.

Así mismo, se aplica a la data los diferentes tipos de análisis de las variables inmersas en la base de datos, comenzando con el Análisis de Componentes principales (ACP).

Para ello segmentamos la población de interés en Condonados y desertados. En la que la población de desertores asciende a 5.105 registros, para este caso se analiza el tipo de Institución de Educación Superior.

**Figura 27**

*Plano factorial desertados ACP.*




*Nota,* Correlación entre grupos. Fuente: Elaboración propia, (2023)

Podemos decir que la inercia es del 59.21%. Se presenta una clusterización en dos grupos representando el género de esta población, teniendo en cuenta que el género es tan representativo, separa las mujeres de los hombres, las demás variables continuas son homogéneas.

En cuanto a la edad, se identifican unos outliers ya que existen personas que superan el promedio de edad. También se evidencian unos pocos outliers en las pruebas saber 11.

**Tabla 10**

*Análisis de Dimensiones Desertados.*

| Variables | Dim.1 | ctr | cos2 | Dim.2 | ctr | cos2 | Dim.3 | ctr | cos2 |
|---|---|---|---|---|---|---|---|---|---|
| **EDAD** | 0,801 | 47,374 | 0,642 | -0,021 | 0,043 | 0,000 | -0,182 | 3,397 | 0,033 |
| **GENERO** | -0,094 | 0,649 | 0,009 | 0,941 | 87,406 | 0,886 | 0,289 | 8,569 | 0,083 |
| **PUNTAJE SABE 11** | -0,784 | 45,339 | 0,614 | -0,236 | 5,507 | 0,056 | 0,130 | 1,742 | 0,017 |
| **PUNTAJE _SISBEN II** | 0,300 | 6,637 | 0,090 | -0,267 | 7,043 | 0,071 | 0,916 | 86,292 | 0,839 |

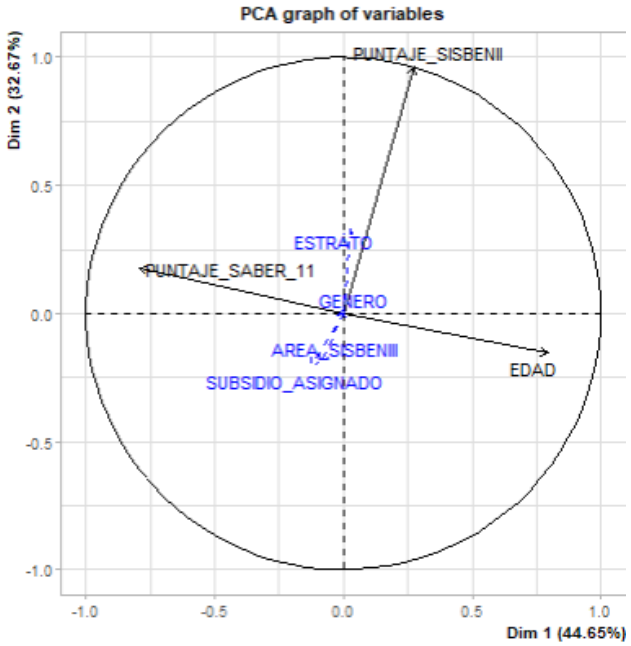
Fuente: Elaboración propia, (2023)

En la dimensión X las variables más representativas son edad y puntaje saber 11. En la dimensión Y la variable más representativa es el género. En una tercera dimensión la variable más representativa es Puntaje Sisbén II.

Se observa en la siguiente imagen que hay una correlación inversa entre edad y puntaje Saber 11, así como una correlación positiva entre Edad y puntaje Sisbén II.

**Figura 28**

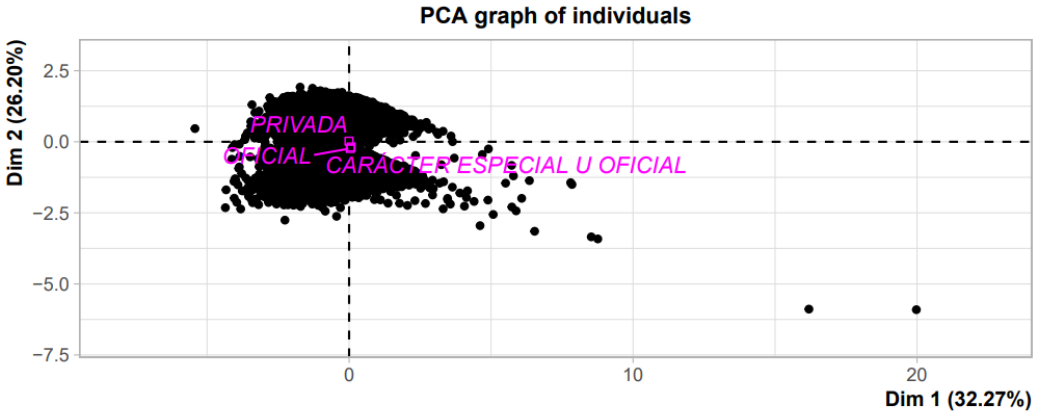
*Grafica de variables desertados - correlaciones ACP.*




*Nota*, Correlaciones entre variables. Fuente: Elaboración propia, (2023)

**Figura 29**

*Plano factorial Condonados ACP.*




*Nota,* Análisis de componentes principales condonados. Fuente: Elaboración propia (2023)

Ahora frente a la población de condonados que corresponde a 11.164, podemos concluir que presentan una inercia de 58.47%, así mismo se evidencia que la población también es homogénea, sin embargo, existen outliers en la variable edad. Por otro lado, no existe mayor diferenciación frente a la selección del tipo de Institución de Educación Superior.

**Tabla 11**

*Análisis de Dimensiones Condonados.*

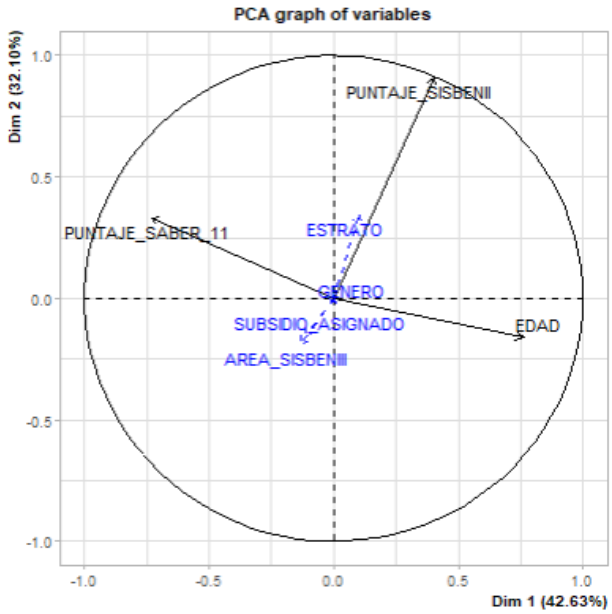
| Variables | Dim.1 | ctr | cos2 | Dim.2 | ctr | cos2 | Dim.3 | ctr | cos2 |
|---|---|---|---|---|---|---|---|---|---|
| **EDAD** | **0,759** | 44,666 | 0,577 | -0,233 | 5,200 | 0,054 | -0,241 | 6,095 | 0,058 |
| **GENERO** | **0,030** | 0,072 | 0,001 | 0,933 | 82,997 | 0,870 | 0,195 | 3,963 | 0,038 |
| **PUNTAJE SABE 11** | **-0,740** | 42,431 | 0,548 | -0,301 | 8,651 | 0,091 | 0,251 | 6,600 | 0,063 |
| **PUNTAJE _SISBEN II** | **0,407** | 12,830 | 0,166 | -0,182 | 3,152 | 0,033 | 0,893 | 83,343 | 0,797 |

Fuente: Elaboración Propia, (2023)

Para la dimensión X se evidencia una alta representación en las variables Edad y Puntaje Saber 11, en la dimensión Y su representación está dada por el género, en una tercera dimensión la variable más representativa es Puntaje Sisbén II.

**Figura 30**

*Grafica de variables condonados- correlaciones ACP.*



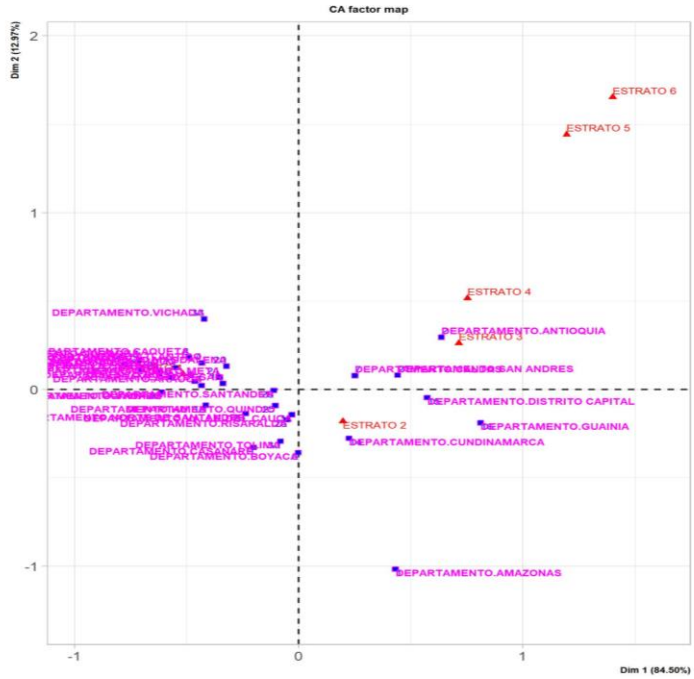

*Nota,* Análisis de componentes principales condonados. Fuente: Elaboración propia, (2023)

En esta gráfica, se evidencia que la variable más representativa es el puntaje sisbenII. La menos representativa es Edad. Existe una mayor correlación positiva entre las variables puntaje Sisbén II y Edad, sin embargo, entre las variables puntaje saber 11 y Edad la correlación es inversa. Con respecto a las variables Edad y Puntaje saber 11 – Puntaje Sisbén II, tiende a ser nula la correlación. Es de señalar que se suprimieron las variables Área Sisbén II, Subsidio Asignado, Género y Estrato por ser variables categóricas.

Ahora, realizaremos un análisis de correspondencia Simple (ACS) tanto de la población condonada como de la población desertada teniendo en cuenta el Estrato y el Departamento de origen de los beneficiarios.

**Figura 31**

*Análisis de Correspondencia Simple Condonados ACS.*

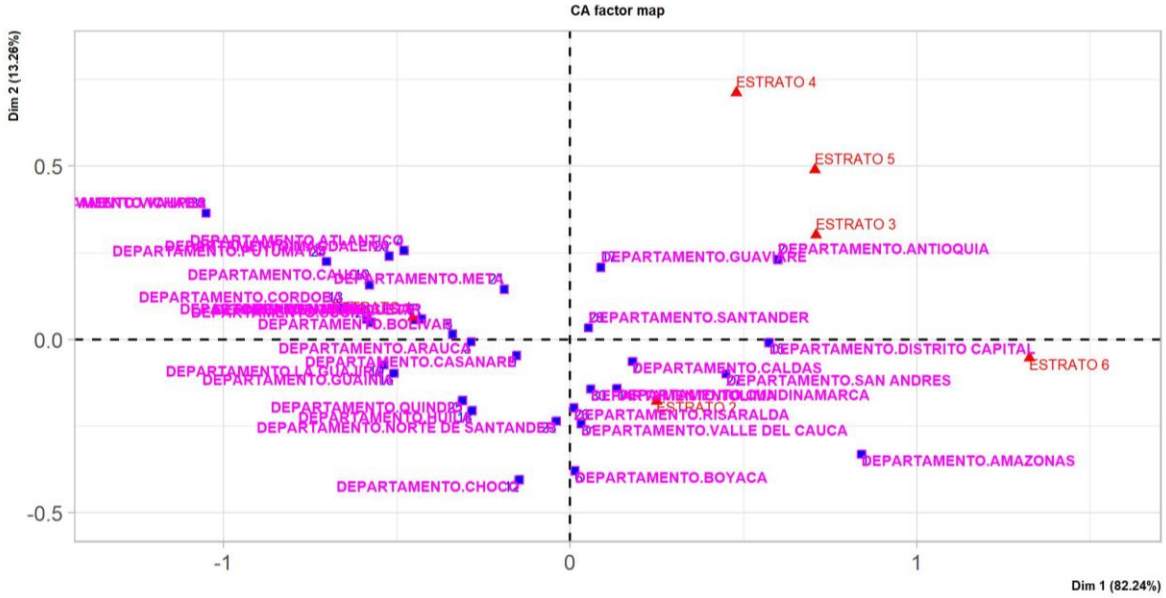



*Nota,* Análisis correspondencias Simples condonados, Fuente: Elaboración propia, (2023)

En la gráfica de análisis de correspondencia simple (ACS) de la población condonados, se evidencia una mayor concentración en el estrato 1 de la mayoría de los departamentos como Vichada, Meta, Santander, Caquetá, Arauca, entre otros. En el estrato 2 tenemos una concentración en el departamento de Cundinamarca, para el estrato 3 se encuentra más representado departamento de Antioquia. Por último, para la ciudad de Bogotá los beneficiarios condonados se encuentran en el estrato 2 y 3. Es importante aclarar que el plano factorial presenta una inercia del 97,47%.

**Figura 32**

*Grafica de variables desertores- correlaciones ACS.*




*Nota*, Análisis correspondencias simples desertores. Fuente: Elaboración propia (2023)

En la gráfica de análisis de correspondencia simple (ACS) de los beneficiarios desertados, se evidencia una mayor concentración en el estrato 1 en la mayoría de los departamentos, como por ejemplo Bolívar, Santander, Córdoba, Arauca, Meta, entre otros.

En el estrato 2 tenemos una concentración en el departamento de Cundinamarca y San Andrés, para el estrato 3 se encuentra más representado departamento de Antioquia. Es importante aclarar que el plano factorial presenta una inercia del 95,5%.

## 7. Experimentación y aplicación del modelo de regresión logística para solución a la problemática.

Una vez realizada la revisión de los modelos predictivos existentes y vistos en la maestría, ver apéndice C, se entiende que la mejor opción es realizar una regresión logística que puede predecir los estudiantes que desertaría o se graduarían.

Sin embargo, al iniciar la experimentación del modelo de regresión logística, encontramos un desbalanceo de los datos, toda vez que la proporción de beneficiarios Inactivos Desertados corresponde al 31.379% frente al 68.621% de los beneficiarios Condonados Graduación, como se evidencia en la tabla 12.

**Tabla 12**

*Desbalanceo de datos*

| **PROGRAMA SPP** | **BENEFICIARIOS** | **PROPORCION** |
|---|---|---|
| CONDONADOS GRADUACION | 11.164 | 68,621% |
| INACTIVOS DESERTORES | 5.105 | 31,379% |
| **TOTAL, BENEFICIARIOS** | **16.269** | **100%** |

*Nota,* Tabla de proporción datos condonados vs desertores. Fuente: Elaboración propia, (2023)

Dado lo anterior, se tiene que hacer el ajuste de los datos desbalanceados usando las técnicas de ajuste de clases.

Para ello se usó el código class_weights, para ajustar los pesos de clase, el cual que se define como un diccionario en el que las claves son las etiquetas de clase y los valores son los pesos correspondientes. La idea detrás del ajuste de pesos de clase es dar más

importancia a las clases minoritarias, para que el modelo las tenga en cuenta al realizar la clasificación. En otras palabras, se trata de aumentar el peso de las observaciones de la clase minoritaria para que tengan más influencia en el ajuste del modelo.

Para tener un acercamiento inicial con el fin de dimensionar el peso de clase específico de la data, se usa el enfoque de equilibrio inverso, que establece el peso de clase como la inversa de la frecuencia de la clase, así:

**Tabla 13**

*Peso Clase*

| BENEFICIARIOS | PROPORCION | PESO CLASE | PESO CLASE |
|---|---|---|---|
| 11.164 | 68,621% | 1/68,621 | 1,46 |
| **5.105** | 31,379% | 1/31,379 | 3,19 |
| **16.269** | **100%** | | |

*Nota*, Calculo de peso de clases de beneficiarios condonados vs desertes. Fuente: Elaboración propia, (2023)

A partir de allí, se realiza las diferentes pruebas que permitan encontrar el mejor equilibrio que permita darle el peso ideal a la clase minoritaria que a su vez es la condición objetivo que se quiere predecir como lo es la deserción. Es por lo que, se define como pesos 0,7286 para la variable mayoritaria y se le asigna un peso de 1,70 a la minoritaria; que tiene una relación similar al equilibrio inverso calculado. Con la anterior definición se continua con el proceso de experimentación.

Para realizar la experimentación del modelo se realizaron trece (13) escenarios en los que se involucraron las diferentes variables que podrían llegar a explicar o apoyar el por

qué los estudiantes desertan de sus programas en educación superior, con los diferentes análisis realizados a través del proyecto se generaron los escenarios que a continuación se exponen.

**Figura 33**

*Variables por escenario*

| | | ESCENARIO | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. | Variables | No. 1 | No. 2 | No. 3 | No. 4 | No. 5 | No. 6 | No. 7 | No. 8 | No. 9 | No. 10 | No. 11 | No. 12 | No. 13 |
| 1 | VERSIÓN | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 2 | ESTRATO | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 3 | EDAD | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 4 | GENERO | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 5 | DEPARTAMENTO_DE_FAMILIA | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 6 | CIUDAD_DE_FAMILIA | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 7 | EDUCACION_PADRE | ✓ | ✓ | ✓ | ✓ | X | X | ✓ | X | ✓ | ✓ | X | X | X |
| 8 | EDUCACION_MADRE | ✓ | ✓ | ✓ | ✓ | X | X | ✓ | X | ✓ | ✓ | X | X | X |
| 9 | OCUPACION_PADRE | ✓ | ✓ | ✓ | ✓ | X | X | ✓ | X | ✓ | ✓ | X | X | X |
| 10 | OCUPACION_MADRE | ✓ | ✓ | ✓ | ✓ | X | X | ✓ | X | ✓ | ✓ | X | X | X |
| 11 | INGRESO_FAMILIAR_MENSUAL | ✓ | ✓ | X | ✓ | X | X | X | X | ✓ | X | X | X | X |
| 12 | DEPARTAMENTO_DE_COLEGIO | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 13 | CIUDAD_DE_COLEGIO | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 14 | COLEGIO | ✓ | X | X | X | X | X | X | X | X | X | X | X | X |
| 15 | COLE_NATURALEZA | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 16 | VALOR_PENSION_COLEGIO | ✓ | ✓ | X | ✓ | X | X | X | X | ✓ | X | X | X | X |
| 17 | COLE_JORNADA | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 18 | PUNTAJE_SABER_11 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 19 | PUNTAJE_SISBENII | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | X | X | X |
| 20 | AREA_SISBENIII | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | X | X | X |
| 21 | SUBSIDIO_ASIGNADO | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ | ✓ | X |
| 22 | ORIGEN_DE_LA_IES | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 23 | IES | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 24 | DEPARTAMENTO_IES | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| 25 | CIUDAD_IES | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | X | X |
| 26 | PROGRAMA | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ |
| | Numero de variables | 26 | 25 | 23 | 22 | 19 | 17 | 22 | 18 | 24 | 21 | 17 | 16 | 15 |

*Nota*, Tabla de variables aplicadas en cada modelo. Fuente: Elaboración propia, (2023).

Una vez definidas las variables a utilizar en los escenarios, se realizó la regresión logística para cada uno de ellos, tomando el 80% de los datos para el entrenamiento y el testeo con el 20% restante, adicionalmente se tuvo en cuenta la situación de tener una variable objetivo (estudiantes desertados) desbalanceada, también se validaron estimadores que ayudaron a tomar la decisión como el AUC, Accuracy, Recall, Precission y F1 Score, para las variables categóricas se realizó la transformación de las mismas en dummies, de tal forma que la regresión logística las pudiera tener en cuenta.

A continuación, se presenta un resumen de los resultados obtenidos con respecto a los evaluadores de los modelos generados y con respecto a nuestra variable objetivo (desertados). Es importante mencionar que el cálculo de los evaluadores (Recall, Precision y F1 Score) se obtuvo con la creación de nuevas matrices de confusión obtenidas de la multiplicación de la matriz obtenida en la base de entrenamiento y los pesos asignados.

**Figura 34**

*Medidas aplicadas a los diferentes escenarios.*

| | No. 1 | No. 2 | No. 3 | No. 4 | No. 5 | No. 6 | No. 7 | No. 8 | No. 9 | No. 10 | No. 11 | No. 12 | No. 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AUC | 0,683 | 0,679 | 0,701 | 0,675 | 0,703 | 0,714 | 0,700 | 0,711 | 0,669 | 0,715 | 0,713 | 0,712 | 0,715 |
| ACURACCY | 0,685 | 0,682 | 0,699 | 0,678 | 0,703 | 0,714 | 0,699 | 0,710 | 0,672 | 0,713 | 0,713 | 0,712 | 0,715 |
| recall (Desertados) | 0,622 | 0,622 | 0,663 | 0,611 | 0,689 | 0,667 | 0,667 | 0,701 | 0,600 | 0,680 | 0,711 | 0,702 | 0,721 |
| recall (Condonados) | 0,742 | 0,736 | 0,739 | 0,740 | 0,718 | 0,734 | 0,734 | 0,721 | 0,738 | 0,750 | 0,715 | 0,723 | 0,710 |
| Precission (Desertados) | 0,689 | 0,684 | 0,734 | 0,683 | 0,728 | 0,731 | 0,731 | 0,734 | 0,678 | 0,754 | 0,724 | 0,727 | 0,723 |
| Precission (Condonados) | 0,681 | 0,680 | 0,669 | 0,674 | 0,678 | 0,670 | 0,670 | 0,687 | 0,668 | 0,675 | 0,702 | 0,697 | 0,707 |
| Fi score (Desertados) | 0,654 | 0,652 | 0,697 | 0,645 | 0,708 | 0,697 | 0,697 | 0,717 | 0,637 | 0,715 | 0,717 | 0,714 | 0,722 |
| Fi score (Condonados) | 0,711 | 0,707 | 0,702 | 0,706 | 0,697 | 0,701 | 0,701 | 0,704 | 0,701 | 0,711 | 0,708 | 0,710 | 0,709 |

*Nota,* Métricas de clasificación aplicadas a modelos. Fuente: Elaboración propia, (2023)

Es de señalar que los resultados individuales como la matriz de confusión se podrán observar en el Apéndice D.

Frente a los resultados presentados, se evidencia que, dentro de los escenarios generados el AUC que evalúa el rendimiento del modelo, arroja un mínimo de 0,669 y un máximo de 0,715; por otra parte, el evaluador Recall que muestra el porcentaje que el modelo identifica correctamente sobre la variable objetivo, que para este caso corresponde a los beneficiarios desertados, teniendo resultados importantes dado que nos muestra un rango entre 0,600 hasta 0,721 siendo este el más alto.

Con respecto al evaluador Precission en cual obtenemos la calidad y que porcentaje de la variable objetivo están en los que predice el modelo, encontramos con un mínimo 0,678 y un máximo de 0,754; para el caso de F1 Score que es la combinación de Precission y Recall se obtuvo un mínimo 0,637 y máximo de 0,722.

Por último, el evaluador Accuracy el cual mide el porcentaje de los aciertos realizados por el modelo, tanto de la variable objetivo como la otra que para este caso es la correspondiente a estudiante graduados y condonados, en los diferentes escenarios se obtuvo un mínimo de 0,672 y máximo de 0,715.

Una vez analizados los resultados y teniendo en cuenta que el objetivo del proyecto empresarial es la identificación de los posibles desertores del programa ser pilo paga, el cual corresponde a una política pública que busca el acceso, permanencia y graduación de la población vulnerable, se define las variables a usar en el modelo y las cuales corresponden al escenario número 13, tomando como primer criterio el AUC y seguido del Recall que entrega un porcentaje de acierto de la variable objetivo del 72%.

**Figura 35**

*Variables del escenario No.13*

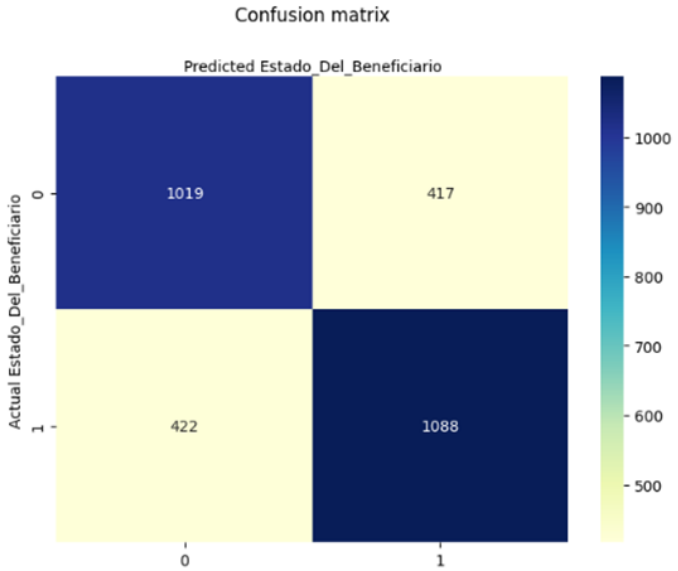
| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | No._Beneficiario | 16269 non-null | int64 |
| 1 | VERSION | 16269 non-null | object |
| 2 | ESTRATO | 16269 non-null | int64 |
| 3 | EDAD | 16269 non-null | int64 |
| 4 | GENERO | 16269 non-null | object |
| 5 | DEPARTAMENTO_DE_FAMILIA | 16269 non-null | object |
| 6 | CIUDAD_DE_FAMILIA | 16269 non-null | object |
| 7 | DEPARTAMENTO_DE_COLEGIO | 16269 non-null | object |
| 8 | CIUDAD_DE_COLEGIO | 16269 non-null | object |
| 9 | COLE_NATURALEZA | 14293 non-null | object |
| 10 | COLE_JORNADA | 14293 non-null | object |
| 11 | PUNTAJE_SABER_11 | 16268 non-null | float64 |
| 12 | ORIGEN_DE_LA_IES | 16269 non-null | object |
| 13 | IES | 16269 non-null | object |
| 14 | DEPARTAMENTO_IES | 16269 non-null | object |
| 15 | PROGRAMA | 16269 non-null | object |

*Nota,* Lista de variables con las cuales se corrió el modelo. Fuente: Elaboración propia, (2023).

En la figura anterior, se relacionan las 15 variables que más inciden en la predicción de deserción de la población del programa ser pilo paga y con las cuales se corrió el modelo y que corresponden al escenario de prueba número trece (13).

**Figura 36**

*Matriz de confusión*




Fuente: Elaboración propia, (2023)

**Figura 37**

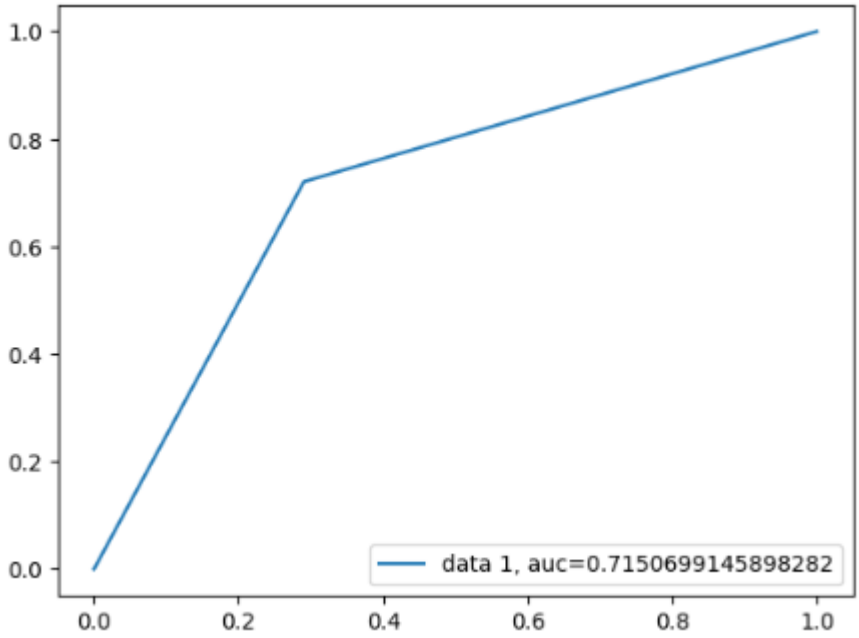
*Resultado de medidas aplicadas al modelo*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.71 | 0.71 | 0.71 | 1436 |
| 1 | 0.72 | 0.72 | 0.72 | 1510 |
| accuracy | | | 0.72 | 2946 |
| macro avg | 0.72 | 0.72 | 0.72 | 2946 |
| weighted avg | 0.72 | 0.72 | 0.72 | 2946 |

Fuente: Elaboración propia, (2023)

**Figura 38**

*AUC curva ROC*




*Nota*, Porcentaje de predicción del modelo. Fuente: Elaboración propia (2023)

Con el modelo definido, se procede a realizar la aplicación del backtesting, en la base de estudiantes que a corte del 30 de junio de 2022 se encontraban estudiando o eran susceptibles de graduación según lo informado por el Ministerio Nacional de Educación Nacional, de acuerdo con la Tabla 3, el número de beneficiarios que se encontraban

estudiando eran 6.749 y número de estudiantes susceptibles de condonación era de 14.878 para un total de 21.627. Una vez realizada la predicción se obtiene los siguientes resultados.

**Tabla 14**

*Predicción*

| Clase | Cantidad |
|---|---|
| **0 - Graduado** | 4.332 |
| **1 - Desertado** | 17.295 |
| **Total, general** | **21.627** |

Fuente: Elaboración propia, (2023)

De acuerdo con la ejecución del modelo, se concluye que de los 21.627 beneficiarios que se encuentran pendiente por graduación o deserción, el 79,97% se predicen por el modelo como desertores (variable objetivo) y por el otro lado, el modelo predice una condonación o graduación del 20,03% de los beneficiarios.

## 8. Evaluación modelo de regresión logística frente a nuevos desertados y graduados

Para la evaluación del modelo adoptado, se busca contrarrestar el resultado del mismo con la situación actual de los beneficiarios, que se hayan graduado o desertado desde el 30 de junio de 2022 a una fecha determinada; así las cosas, se recibió por parte del Ministerio de Educación Nacional, la información de los beneficiarios que a corte del 16 de abril de 2023 cambiaron su estado por graduado o desertado, que para este caso corresponde a un total de 2.897 beneficiarios, que están representados por 237 beneficiarios desertados y 2.660 beneficiarios graduados o condonados. A partir de esta información se hace la validación, donde se obtiene los siguientes resultados.

**Figura 39**

*Resultados predicción nueva población.*

| | | CONDONADOS GRADUACION | INACTIVOS DESERTORES | TOTAL |
|---|---|---|---|---|
| REALIDAD | CONDONADOS GRADUACION | 1684 | 976 | 2660 |
| | INACTIVOS DESERTORES | 15 | 222 | 237 |
| | TOTAL | 1699 | 1198 | 2897 |

*Nota,* Resultados Backtesting. Fuente: Elaboración propia, (2023).

Se puede identificar que el modelo predice el 93,67% de los casos donde el beneficiario deserta, que es la variable objetivo, siendo un valor relevante para la atención a la problemática que se viene presentando en dicha población, en donde se podrá realizar una validación de los estudiantes que el modelo pronostica como desertores para adelantar

un proceso de acompañamiento, con el fin de cumplir la naturaleza del programa creado desde el Gobierno Nacional, que busca que las personas con menores recursos y resultados académicos sobresalientes, tengan la posibilidad de obtener un título de educación superior para ayudar a mejorar su calidad de vida.

Aunque el modelo presenta un número considerable de casos que se predicen como desertados siendo graduados o condonados en la realidad, el interés de este proyecto es generar alertas para realizar acompañamientos a los posibles desertores y brindar apoyo y ayuda necesaria con el fin de que logren la graduación y cambiar los proyectos de vida, este costo de acompañamiento sería menor, comparado con la perdida de los recursos ya invertidos y el costo social que se puede dar a una persona que no accede, permanece y se gradúa en educación superior.

# 9. Conclusiones

En cumplimiento de los objetivos propuestos, se puede concluir que en el desarrollo de este proyecto empresarial se usaron diversas herramientas analíticas, para lograr la identificación de los beneficiarios desertores del programa ser pilo paga, esto con el fin de que se desarrollen las estrategias necesarias para realizar acompañamiento a estos jóvenes y así fortalecer los procesos de permanencia y graduación en la educación superior del país.

Se determinó que el modelo ideal a usar para la predicción de la deserción académica en la educación superior de la población vulnerable del programa ser pilo paga, es el modelo de regresión logística, así mismo, se lograron identificar las variables socioeconómicas y académicas que influyen en mayor proporción en la deserción de los jóvenes.

Se identificaron las 15 variables que al ejecutar el modelo infieren en la deserción académica de los jóvenes del programa ser Pilo Paga, las cuales son: Versión, Estrato, Edad, Género, Departamento de Familia, Ciudad de familia, Departamento del colegio, Ciudad del colegio, Naturaleza del colegio, Jornada del colegio, Puntaje saber 11, Origen de la IES, IES, Departamento IES y Programa.

Dentro de las medidas usadas para evaluar el modelo ejecutado, se realizó el cálculo de AUC, Accuracy, recall, precision, F1 – Score.

Al replicar el modelo a la población que se encontraba en curso de sus estudios, se determinó un 93% de efectividad en la predicción de jóvenes que van a desertar a futuro, por otro lado, se presenta un 63% de efectividad en la predicción de graduados. Dado esto, se considera que el costo de hacer un acompañamiento a los jóvenes con predicción de deserción sería menor a la pérdida económica y social que se puede presentar en un verdadero caso de ser desertor.

## 10. Referencias Bibliográficas

Betancourth Sánchez, L. J., & Cuesta, J. (2016). *Orientación vocacional y profesional en la juventud colombiana* (edsair.od.....2802..1df0f1b7ada109e076af7519a26bf99b). OpenAIRE. https://explore.openaire.eu/search/publication?articleId=od_____2802::1df0f1b7ada109e076af7519a26bf99b

Martínez Ipuz, J. A., Flórez, K., García, H., Gómez, V. M., González Jiménez, D. A., Bravo Castillo, M., Cuervo, F., Walker Forero, C., Gandini Price, A., Cardona Abrego, M., Garzón Rayo, O., Costala, G., Miranda, N., & Andrade, M. I. (2009). *Retraimiento poblacional en educación superior: Ingreso, mortalidad académica y deserción*. Universidad de San Buenaventura - Cali.

Ministerio de Educación Nacional de Colombia. (2017a, febrero 7). *Deserción Escolar*. https://www.mineducacion.gov.co/portal/secciones/Glosario/82745:DESERCION-ESCOLAR

Ministerio de Educación Nacional de Colombia. (2017b, febrero 27). *Población Vulnerable*. https://www.mineducacion.gov.co/portal/secciones/Glosario/82770:POBLACION-VULNERABLE

Mora Cortés, A. F. (2016). *La seudorrevolución educativa: Desigualdades, capitalismo y control en la educación superior en Colombia* (edselb.123305). eLibro. https://search.ebscohost.com/login.aspx?direct=true&db=edselb&AN=edselb.123305&site=eds-live

Palmer Pol, A., & Montaño Moreno, J. J. (1999). ¿Qué son las redes neuronales artificiales? Aplicaciones realizadas en el ámbito de las adicciones. *Adicciones: Revista de Socidrogalcohol*, *11*(3), 243-243-255.

República de Colombia. (1991). *Constitución política de Colombia*. Legis.

# Apéndice

## Apéndice A Matriz de riesgos

## Figura A1

*Matriz de Riesgos*

| Nombre del Proceso | Nombre del Activo | Tipo de Activo | Impacto | Amenaza | Vulnerabilidad | Probabilidad del Escenario | Riesgo Inicial | Estado de los Controles Actuales | Efectividad del Control | Riesgo Residual | Aceptación | Tratamiento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PROYECTO EMPRESARIAL | BASE DE DATOS SPP | INFORMACIÓN DIGITAL | 5,0 | Robo | Falta de controles de almacenamiento y resguardo | 3 | 15 | Se cuenta con Back Up de la información. | 4 | 3,75 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | BASE DE DATOS SPP | INFORMACIÓN DIGITAL | 5,0 | Alteración o Eliminación | Falta de políticas, normas o procedimientos | 2 | 10 | | 1 | 10 | ✗ | Generar Contraseña de acceso Dejar archivo de solo lectura |
| PROYECTO EMPRESARIAL | PUNTAJE PRUEBAS SABER 11 | INFORMACIÓN DIGITAL | 5,0 | Alteración o Eliminación | Falta controles de gestión de cambio | 2 | 10 | Se cuenta con Back Up de la información. | 3 | 3,333 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | BIBLIOGRAFIA | INFORMACIÓN DIGITAL | 1,0 | Alteración o Eliminación | Ausencia de copias de respaldo | 3 | 3 | Se cuenta con respaldo bibliografico en aplicativo zotero | 4 | 0,75 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | MENSAJES DE CORREOS ELECTRONICO | INFORMACIÓN DIGITAL | 5,0 | Fuga de Información | Protección inadecuada de trafico sensible | 3 | 15 | | 1 | 15 | ✗ | Se usara aplicativo mailvelope para el envio de archivos por correo electronico |
| PROYECTO EMPRESARIAL | MENSAJES DE CORREOS ELECTRONICO | INFORMACIÓN DIGITAL | 4,0 | Alteración o Eliminación | Empleado Insatisfecho | 1 | 4 | | 1 | 4 | ✓ | |
| PROYECTO EMPRESARIAL | MENSAJES DE CORREOS ELECTRONICO | INFORMACIÓN DIGITAL | 5,0 | Daño accidental | Falta de control de documentos | 2 | 10 | El correo almacena borradores, mensajes eliminados, mensajes enviados | 3 | 3,333 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | COMPUTADORES PERSONALES | HADWARE | 4,0 | Robo | Falta de controles de activos | 3 | 12 | Se cuenta con Back Up de la información tanto en la nube como en correos electronicos | 4 | 3 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | COMPUTADORES PERSONALES | HADWARE | 4,0 | Sobrecarga de trafico | Capacidad inadecuada de red | 1 | 4 | | 1 | 4 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | DERECHO DE PETICION | INFORMACION FISICA | 2,5 | Daño accidental | Falta de conciencia en seguridad | 3 | 7,5 | Se cuenta con radicado por parte de la entidad receptora | 3 | 2,5 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | RESPUESTA DE DERECHO DE PETICION | INFORMACION FISICA | 5,0 | Denegación de Servicio | Indisponibilidad de datos de respaldo | 2 | 10 | reorientacion del product backlog | 3 | 3,333 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | ACUERDO DE CONFIDENCIALIDAD | INFORMACION FISICA | 4,0 | Alteración o Eliminación | Ausencia de copias de respaldo | 3 | 12 | se cuenta con copia del documento fisico y digital | 4 | 3 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | CARTA DE COMPROMISO | INFORMACION FISICA | 4,0 | Alteración o Eliminación | Ausencia de copias de respaldo | 3 | 12 | se cuenta con copia del documento fisico y digital | 4 | 3 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | GOOGLE DRIVE | SERVICIO | 2,0 | Denegación de Servicio | Falta de dispositivos de contuidad electrica | 1 | 2 | se cuenta con backup de la información | 4 | 0,5 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | BIG QUERY | SERVICIO | 2,0 | Denegación de Servicio | Falta de dispositivos de contuidad electrica | 2 | 4 | | 1 | 4 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | DEVELOPERS (JAIME Y PILAR) | RECURSO HUMANO | 5,0 | Fuga de Información | Empleado Insatisfecho | 2 | 10 | se cuenta con backup de la información | 3 | 3,333 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | DIRECTOR DE PROYECTO | RECURSO HUMANO | 3,0 | Denegación de Servicio | Falta de planes de continuidad | 2 | 6 | Existe carta de compromiso | 3 | 2 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | PRODUCT OWNER | RECURSO HUMANO | 4,0 | Falla | Sobrecarga de labores | 3 | 12 | Programacion con antelacion de reuniones Apoyos con encargados del negocio | 3 | 4 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | SCRIPT DEL MODELO | SOFTWARE | 5,0 | Infección / Software Malicioso | Falta de protección contra código malicioso | 2 | 10 | Se cuenta con antivirus en los equipos | 3 | 3,333 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | POWER BI | SOFTWARE | 3,5 | Falla | Configuracion inadecuada | 2 | 7 | Guias de instalacion y uso del software | 3 | 2,333 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | DATA STUDIO | SOFTWARE | 3,5 | Indisponibilidad | Caida de servidor | 2 | 7 | Backup del Script | 3 | 2,333 | ✓ | RIESGO ACEPTADO |
| PROYECTO EMPRESARIAL | RSTUDIO | SOFTWARE | 3,5 | Alteración o Eliminación | Falta controles de gestión de cambio | 2 | 7 | Acceso a las paginas oficiales para descarga de instaladores en caso eliminacion. | 3 | 2,333 | ✓ | RIESGO ACEPTADO |

*Nota.* Muestra matriz de riesgos al inicio del proyecto empresarial. Fuente: Elaboración Propia, (2023)

## Apéndice B Cronograma

Para el desarrollo del proyecto analítico que se requiere adelantar según la metodología usada y posterior generación del modelo, se desarrolló un cronograma a partir de las metodologías agiles vistas en la maestría, en las cuales se plasma las etapas definidas, pero no secuenciales como se maneja en la metodología tradicional, sino que, en caso de necesitar retornar a otra etapa, de acuerdo con lo definido en cada review de los sprint como se evidencia en la Figura 1. Sprint del proyecto.

**Figura B1**

*Sprint del proyecto*


*Nota.* Muestra el diseño visual de un sprint del proyecto empresarial. Fuente: Elaboración propia, (2023)

A continuación, se muestra el cronograma de trabajo, con sus diferentes sprint, tareas, entregables y fechas estimadas.

**Tabla B1**

*Cronograma*

| Nombre de tarea | Duración | Comienzo | Fin | Entregables |
|---|---|---|---|---|
| **Diseño** | **84 días** | **2/04/2022** | **25/06/2022** | Ante Proyecto |
| Definición de alcance (Objetivos y Justificación) | 41 días | 2/04/2022 | 12/05/2022 | |
| Definición de Cronograma | 8 días | 13/05/2022 | 20/05/2022 | |
| Identificación Fuentes de Información | 56 días | 2/04/2022 | 28/05/2022 | |
| Metodología | 18 días | 30/05/2022 | 22/06/2022 | |
| Estudio del Arte (Avance) Marco Teórico | 18 días | 30/05/2022 | 22/06/2022 | |
| Bibliografía | 18 días | 30/05/2022 | 22/06/2022 | |
| Review | 1 día | 22/06/2022 | 22/06/2022 | |
| Retrospectiva | 1 día | 25/06/2022 | 25/06/2022 | |
| **Valoración de Data** | **26 días** | **30/05/2022** | **25/06/2022** | 1. Base definitiva a trabajar<br>2. Tablero de visualización (preliminar) |
| Recolección de data | 18 días | 30/05/2022 | 22/06/2022 | |
| Análisis y de depuración de las bases de datos | 9 días | 10/06/2022 | 22/06/2022 | |
| Estadística Descriptiva | 9 días | 10/06/2022 | 22/06/2022 | |
| Review (Contemplar si es necesario devolverse a una etapa) | 1 día | 22/06/2022 | 22/06/2022 | |
| Retrospectiva | 1 día | 25/06/2022 | 25/06/2022 | |
| **Modelación** | **306 días** | **26/06/2022** | **28/04/2023** | 1. Salidas de correlación, tendencias, etc.<br>2. Resultados de diferentes herramientas de predicción<br>3. Comparativo de significancias, efectividad de los |
| Análisis y minería de datos | 32 días | 26/06/2022 | 6/08/2022 | |
| Comprensión de tendencias, identificación de correlaciones y variaciones | 90 días | 8/08/2022 | 9/12/2022 | |
| Experimentación con las herramientas de predicción | 70 días | 01/12/2022 | 9/02/2023 | |

| Nombre de tarea | Duración | Comienzo | Fin | Entregables |
|---|---|---|---|---|
| Validación de significancias, % de efectividad y Predicción | 90 días | 09/12/2022 | 9/03/2023 | modelos<br>4. Conclusiones y toma del modelo ideal |
| Determinación de modelo a utilizar | 16 días | 15/03/2023 | 31/03/2023 | |
| Review (Contemplar si es necesario devolverse a una etapa) | 1 día | 10/04/2023 | 11/04/2023 | |
| Retrospectiva | 1 día | 12/04/2023 | 12/04/2023 | |
| **Implementación y testeo** | 20 días | 4/04/2023 | 24/04/2023 | 1. Resultados de predicción<br>2. Informe y conclusiones |
| Ejecución del modelo | 20 días | 4/04/2023 | 24/04/2023 | |
| Prueba con los demás beneficiarios | 8 días | 20/04/2023 | 28/04/2023 | |
| Evaluación estadística de los resultados | 5 días | 20/04/2023 | 25/04/2023 | |
| Review (Contemplar si es necesario devolverse a una etapa) | 1 día | 26/04/2023 | 26/04/2023 | |
| Retrospectiva | 1 día | 27/04/2023 | 27/04/2023 | |
| **Evaluación del Modelo** | **9 días** | **20/04/2023** | **29/04/2023** | 1. Presentación de informe con indicador generado: número de aciertos del modelo / No. de beneficiarios reales graduados o desertados |
| Identificar los beneficiarios que se encontraban en estado activo y que ya presentan graduación o deserción a corte de 16 de abril de 2023 | 7 días | 20/04/2023 | 27/04/2023 | |
| Contrastar la predicción del modelo versus la situación real de los beneficiarios | 4 días | 24/04/2023 | 28/04/2023 | |
| Creación del indicador | 1 días | 27/04/2023 | 28/04/2023 | |
| Review (Contemplar si es necesario devolverse a una etapa) | 1 días | 28/04/2023 | 29/04/2023 | |
| Retrospectiva | 1 día | 28/04/2023 | 28/04/2023 | |
| **Visualización** | **25 días** | **02/05/2023** | **27/05/2023** | 1. Tablero de visualización de resultados del modelo<br>2. Informe |
| Data Storytelling | 8 días | 02/05/2023 | 10/05/2023 | |
| Elaboración y presentación de Conclusiones | 7 días | 20/05/2023 | 27/05/2023 | |

| Nombre de tarea | Duración | Comienzo | Fin | Entregables |
|---|---|---|---|---|
| Review | 1 día | 22/05/2023 | 22/05/2023 | |
| Retrospectiva – Cierre | 1 día | 23/05/2023 | 23/05/2023 | |

*Nota.* Se registra cronograma de actividades realizadas en proyecto empresarial. Fuente: Elaboración Propia, (2023)

**Apéndice C Descripción de las alternativas, estrategias y/o acciones**

Dentro de la cantidad de modelos predictivos vistos en la maestría, hemos identificado el modelo ideal a utilizar para el proyecto empresarial, sin embargo, antes queremos citar los demás modelos que hemos conocido y explorado en algunos talleres prácticos de la maestría. Iniciaremos por los modelos basados en redes neuronales, que según Palmer y Montaño son: algoritmos que pretenden emular este comportamiento dando lugar a un proceso de aprendizaje automático (1999), que para nuestro proyecto empresarial no aplica.

Pasando ahora a los modelos no supervisados encontramos la posibilidad de realizar clasificación de la población basada en las variables que consideremos relevantes para la construcción de los grupos que deseemos tengan alguna caracterización, es decir, que el comportamiento o características de un grupo sean similares pero tengan diferencia con los demás grupos de tal forma que esto nos sirva para la toma de decisión con respecto a tomar acciones en cada uno de los grupos formados; este tipo de modelos se encuentran desarrollados cuando no se tiene una variable objetivo claramente definida.

En nuestro caso, para el proyecto empresarial deseamos predecir la población que podría desertar u obtener el grado, utilizando las variables socioeconómicas, académicas y geográficas como explicación dentro del modelo, es decir, tenemos una variable objetivo-definida que es deserción o graduación de un programa académico en la población

vulnerable. En este caso no sería viable la utilización de dichos modelos dado el problema a solucionar.

Por último, explorando los modelos supervisados donde su naturaleza es tener una variable objetivo, encontramos que por medio de la utilización de variables independientes podemos llegar a explicar por qué una variable se comporta de alguna manera definida, es por ello que para nuestro caso, al explicar la deserción o graduación de los beneficiarios del programa SPP, el gran reto es encontrar las variables adecuadas y que por el criterio experto sean congruentes con el comportamiento de la variable dependiente.

Con la exploración de estos modelos identificamos que la regresión logística se adecua a la necesidad del proyecto, como se ha mencionado nos permite explicar una variable dependiente (deserción) con las variables independientes, a su vez este modelo nos permite tener un resultado binario que, para el caso, indicará si deserta o se gradúa el beneficiario, teniendo en cuenta que se debe evitar la colinealidad de las variables independientes con el fin de que el modelo sea representativo.

## Apéndice D Diferentes escenarios de aplicación del modelo

**Figura D1**

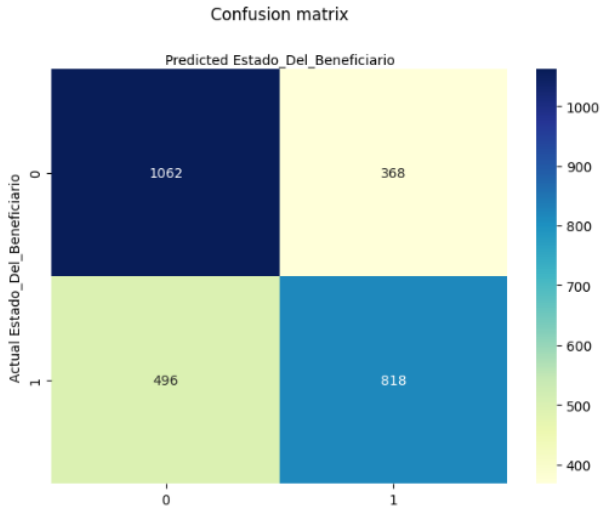
*Variables escenario 1*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 28 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   No._Beneficiario          16269 non-null  int64
 1   VERSION                   16269 non-null  object
 2   ESTRATO                   16269 non-null  int64
 3   EDAD                      16269 non-null  int64
 4   GENERO                    16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA   16269 non-null  object
 6   CIUDAD_DE_FAMILIA         16269 non-null  object
 7   EDUCACION_PADRE           14300 non-null  object
 8   EDUCACION_MADRE           14300 non-null  object
 9   OCUPACION_PADRE           14299 non-null  object
 10  OCUPACION_MADRE           14297 non-null  object
 11  INGRESO_FAMILIAR_MENSUAL  13794 non-null  object
 12  DEPARTAMENTO_DE_COLEGIO   16269 non-null  object
 13  CIUDAD_DE_COLEGIO         16269 non-null  object
 14  COLEGIO                   16268 non-null  object
 15  COLE_NATURALEZA           14293 non-null  object
 16  VALOR_PENSION_COLEGIO     13745 non-null  object
 17  COLE_JORNADA              14293 non-null  object
 18  PUNTAJE_SABER_11          16268 non-null  float64
 19  PUNTAJE_SISBENII          16190 non-null  float64
 20  AREA_SISBENIII            16190 non-null  object
 21  SUBSIDIO_ASIGNADO         16269 non-null  float64
 22  ORIGEN_DE_LA_IES          16269 non-null  object
 23  IES                       16269 non-null  object
 24  DEPARTAMENTO_IES          16269 non-null  object
 25  CIUDAD_IES                16269 non-null  object
 26  PROGRAMA                  16269 non-null  object
 27  ESTADO_DEL_BENEFICIARIO   16269 non-null  int64
dtypes: float64(3), int64(4), object(21)
memory usage: 3.5+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D2**

*Matriz de confusión escenario 1*




Fuente; Elaboración propia, (2023)

**Figura D3**

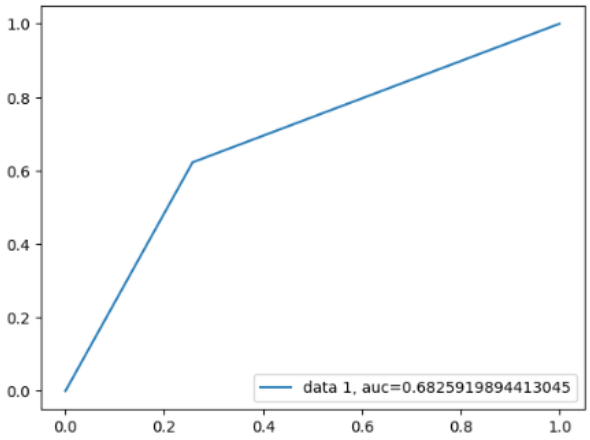
*Medidas escenario 1*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.74 | 0.71 | 1430 |
| 1 | 0.69 | 0.62 | 0.65 | 1314 |
| accuracy | | | 0.69 | 2744 |
| macro avg | 0.69 | 0.68 | 0.68 | 2744 |
| weighted avg | 0.69 | 0.69 | 0.68 | 2744 |

Fuente; Elaboración propia, (2023)

**Figura D4**

*Curva Roc escenario 1*




Fuente; Elaboración propia, (2023)

**Figura D5**

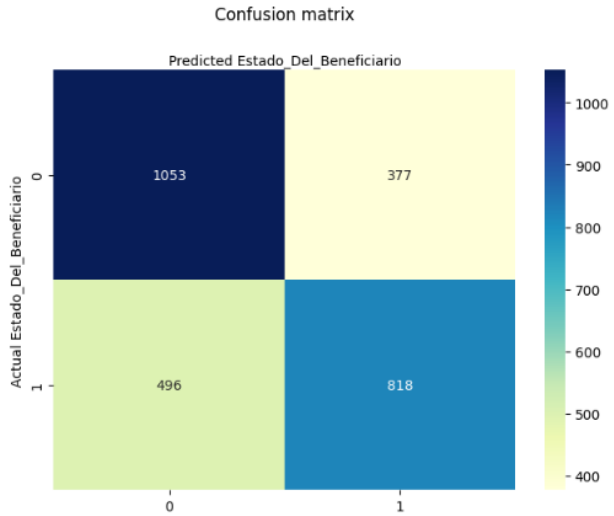
*Variables escenario 2*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 27 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   No._Beneficiario          16269 non-null  int64
 1   VERSION                   16269 non-null  object
 2   ESTRATO                   16269 non-null  int64
 3   EDAD                      16269 non-null  int64
 4   GENERO                    16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA   16269 non-null  object
 6   CIUDAD_DE_FAMILIA         16269 non-null  object
 7   EDUCACION_PADRE           14300 non-null  object
 8   EDUCACION_MADRE           14300 non-null  object
 9   OCUPACION_PADRE           14299 non-null  object
 10  OCUPACION_MADRE           14297 non-null  object
 11  INGRESO_FAMILIAR_MENSUAL  13794 non-null  object
 12  DEPARTAMENTO_DE_COLEGIO   16269 non-null  object
 13  CIUDAD_DE_COLEGIO         16269 non-null  object
 14  COLE_NATURALEZA           14293 non-null  object
 15  VALOR_PENSION_COLEGIO     13745 non-null  object
 16  COLE_JORNADA              14293 non-null  object
 17  PUNTAJE_SABER_11          16268 non-null  float64
 18  PUNTAJE_SISBENII          16190 non-null  float64
 19  AREA_SISBENIII            16190 non-null  object
 20  SUBSIDIO_ASIGNADO         16269 non-null  float64
 21  ORIGEN_DE_LA_IES          16269 non-null  object
 22  IES                       16269 non-null  object
 23  DEPARTAMENTO_IES          16269 non-null  object
 24  CIUDAD_IES                16269 non-null  object
 25  PROGRAMA                  16269 non-null  object
 26  ESTADO_DEL_BENEFICIARIO   16269 non-null  int64
dtypes: float64(3), int64(4), object(20)
memory usage: 3.4+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D6**

*Matriz de confusión escenario 2*




Fuente; Elaboración propia, (2023)

**Figura D7**

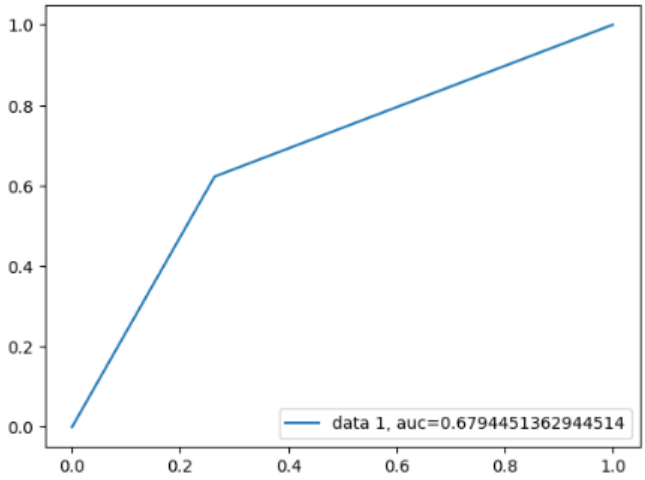
*Medidas escenario 2*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.74 | 0.71 | 1430 |
| 1 | 0.68 | 0.62 | 0.65 | 1314 |
| accuracy | | | 0.68 | 2744 |
| macro avg | 0.68 | 0.68 | 0.68 | 2744 |
| weighted avg | 0.68 | 0.68 | 0.68 | 2744 |

Fuente; Elaboración propia, (2023)

***Figura D8***

*Curva ROC escenario 2*




Fuente; Elaboración propia, (2023)

**Figura D9**

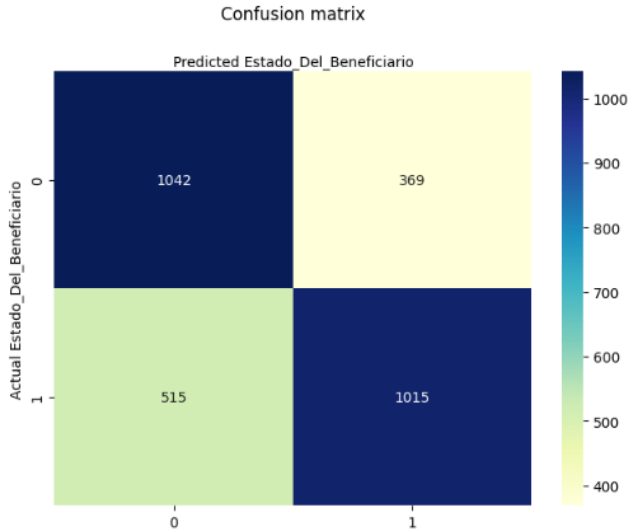
*Variables escenario 3*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 25 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   No._Beneficiario         16269 non-null  int64
 1   VERSION                  16269 non-null  object
 2   ESTRATO                  16269 non-null  int64
 3   EDAD                     16269 non-null  int64
 4   GENERO                   16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA  16269 non-null  object
 6   CIUDAD_DE_FAMILIA        16269 non-null  object
 7   EDUCACION_PADRE          14300 non-null  object
 8   EDUCACION_MADRE          14300 non-null  object
 9   OCUPACION_PADRE          14299 non-null  object
 10  OCUPACION_MADRE          14297 non-null  object
 11  DEPARTAMENTO_DE_COLEGIO  16269 non-null  object
 12  CIUDAD_DE_COLEGIO        16269 non-null  object
 13  COLE_NATURALEZA          14293 non-null  object
 14  COLE_JORNADA             14293 non-null  object
 15  PUNTAJE_SABER_11         16268 non-null  float64
 16  PUNTAJE_SISBENII         16190 non-null  float64
 17  AREA_SISBENIII           16190 non-null  object
 18  SUBSIDIO_ASIGNADO        16269 non-null  float64
 19  ORIGEN_DE_LA_IES         16269 non-null  object
 20  IES                      16269 non-null  object
 21  DEPARTAMENTO_IES         16269 non-null  object
 22  CIUDAD_IES               16269 non-null  object
 23  PROGRAMA                 16269 non-null  object
 24  ESTADO_DEL_BENEFICIARIO  16269 non-null  int64
dtypes: float64(3), int64(4), object(18)
memory usage: 3.1+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D10**

*Matriz de confusión escenario 3*




Fuente; Elaboración propia, (2023)

**Figura D11**

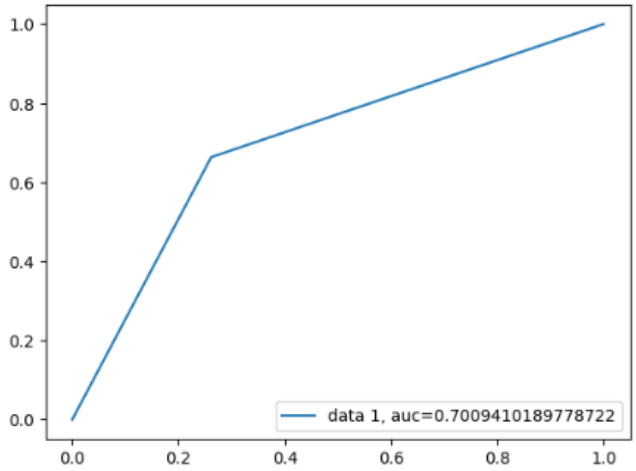
*Medidas escenario 3*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.74 | 0.70 | 1411 |
| 1 | 0.73 | 0.66 | 0.70 | 1530 |
| accuracy | | | 0.70 | 2941 |
| macro avg | 0.70 | 0.70 | 0.70 | 2941 |
| weighted avg | 0.70 | 0.70 | 0.70 | 2941 |

Fuente; Elaboración propia, (2023)

**Figura D12**

*Curva ROC escenario 3*




Fuente; Elaboración propia, (2023)

**Figura D13**

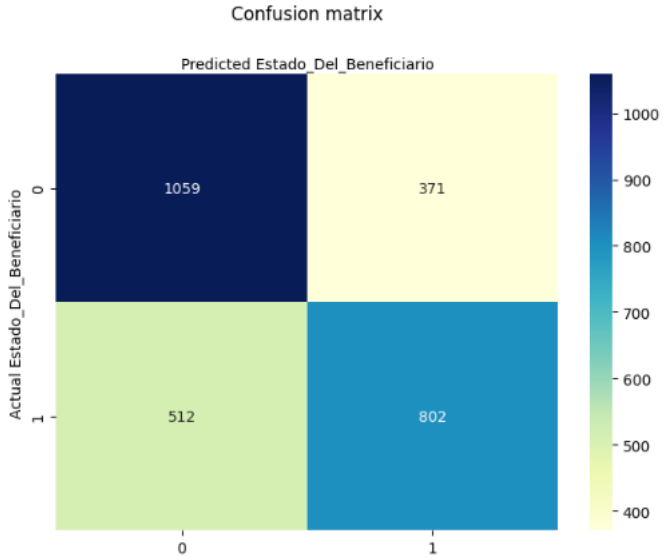
*Variables escenario 4*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 24 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   No._Beneficiario          16269 non-null  int64
 1   VERSION                   16269 non-null  object
 2   ESTRATO                   16269 non-null  int64
 3   EDAD                      16269 non-null  int64
 4   GENERO                    16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA   16269 non-null  object
 6   EDUCACION_PADRE           14300 non-null  object
 7   EDUCACION_MADRE           14300 non-null  object
 8   OCUPACION_PADRE           14299 non-null  object
 9   OCUPACION_MADRE           14297 non-null  object
 10  INGRESO_FAMILIAR_MENSUAL  13794 non-null  object
 11  DEPARTAMENTO_DE_COLEGIO   16269 non-null  object
 12  COLE_NATURALEZA           14293 non-null  object
 13  VALOR_PENSION_COLEGIO     13745 non-null  object
 14  COLE_JORNADA              14293 non-null  object
 15  PUNTAJE_SABER_11          16268 non-null  float64
 16  PUNTAJE_SISBENII          16190 non-null  float64
 17  AREA_SISBENIII            16190 non-null  object
 18  SUBSIDIO_ASIGNADO         16269 non-null  float64
 19  ORIGEN_DE_LA_IES          16269 non-null  object
 20  IES                       16269 non-null  object
 21  DEPARTAMENTO_IES          16269 non-null  object
 22  PROGRAMA                  16269 non-null  object
 23  ESTADO_DEL_BENEFICIARIO   16269 non-null  int64
dtypes: float64(3), int64(4), object(17)
memory usage: 3.0+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D14**

*Matriz de confusión escenario 4*




Fuente; Elaboración propia, (2023)

**Figura D15**

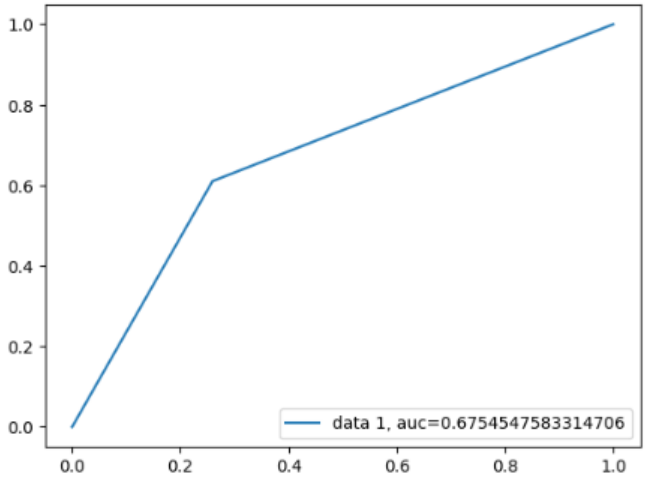
*Medidas escenario 4*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.74 | 0.71 | 1430 |
| 1 | 0.68 | 0.61 | 0.64 | 1314 |
| accuracy | | | 0.68 | 2744 |
| macro avg | 0.68 | 0.68 | 0.68 | 2744 |
| weighted avg | 0.68 | 0.68 | 0.68 | 2744 |

Fuente; Elaboración propia, (2023)

**Figura D16**

*Curva ROC escenario 4*




Fuente; Elaboración propia, (2023)

**Figura D17**

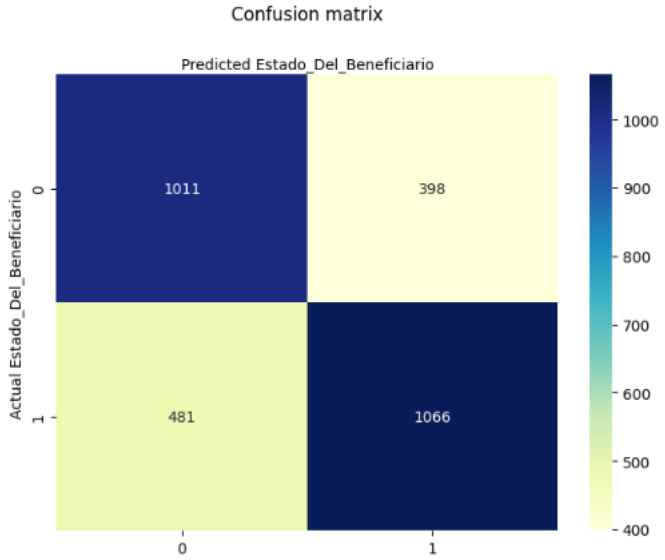
*Variables escenario 5*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 21 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   No._Beneficiario         16269 non-null  int64
 1   VERSION                  16269 non-null  object
 2   ESTRATO                  16269 non-null  int64
 3   EDAD                     16269 non-null  int64
 4   GENERO                   16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA  16269 non-null  object
 6   CIUDAD_DE_FAMILIA        16269 non-null  object
 7   DEPARTAMENTO_DE_COLEGIO  16269 non-null  object
 8   CIUDAD_DE_COLEGIO        16269 non-null  object
 9   COLE_NATURALEZA          14293 non-null  object
 10  COLE_JORNADA             14293 non-null  object
 11  PUNTAJE_SABER_11         16268 non-null  float64
 12  PUNTAJE_SISBENII         16190 non-null  float64
 13  AREA_SISBENIII           16190 non-null  object
 14  SUBSIDIO_ASIGNADO        16269 non-null  float64
 15  ORIGEN_DE_LA_IES         16269 non-null  object
 16  IES                      16269 non-null  object
 17  DEPARTAMENTO_IES         16269 non-null  object
 18  CIUDAD_IES               16269 non-null  object
 19  PROGRAMA                 16269 non-null  object
 20  ESTADO_DEL_BENEFICIARIO  16269 non-null  int64
dtypes: float64(3), int64(4), object(14)
memory usage: 2.6+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D18**

*Matriz de confusión escenario 5*




Fuente; Elaboración propia, (2023)

**Figura D19**

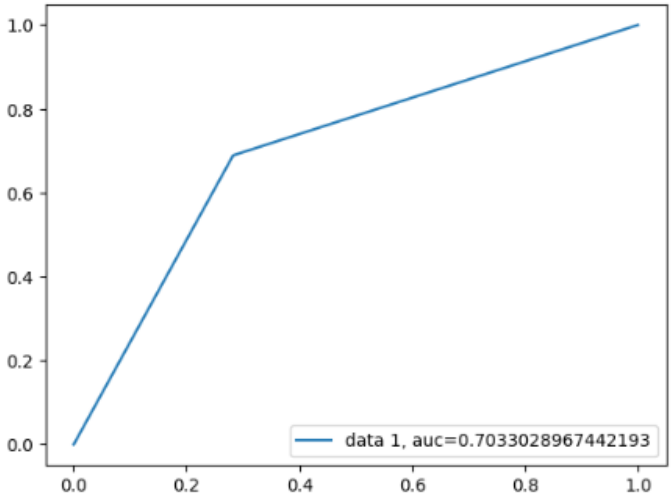
*Medidas escenario 5*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.72 | 0.70 | 1409 |
| 1 | 0.73 | 0.69 | 0.71 | 1547 |
| accuracy | | | 0.70 | 2956 |
| macro avg | 0.70 | 0.70 | 0.70 | 2956 |
| weighted avg | 0.70 | 0.70 | 0.70 | 2956 |

Fuente; Elaboración propia, (2023)

**Figura D20**

*Curva ROC escenario 5*




Fuente; Elaboración propia, (2023)

**Figura D21**

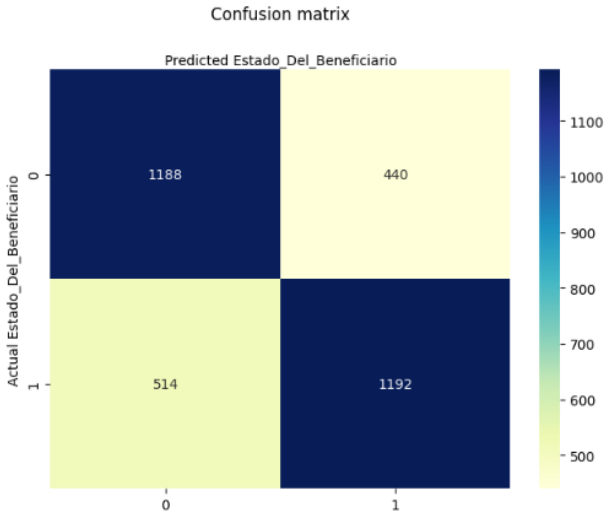
*Variables escenario 6*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 19 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   No._Beneficiario         16269 non-null  int64
 1   VERSION                  16269 non-null  object
 2   ESTRATO                  16269 non-null  int64
 3   EDAD                     16269 non-null  int64
 4   GENERO                   16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA  16269 non-null  object
 6   CIUDAD_DE_FAMILIA        16269 non-null  object
 7   DEPARTAMENTO_DE_COLEGIO  16269 non-null  object
 8   CIUDAD_DE_COLEGIO        16269 non-null  object
 9   PUNTAJE_SABER_11         16268 non-null  float64
 10  PUNTAJE_SISBENII         16190 non-null  float64
 11  AREA_SISBENIII           16190 non-null  object
 12  SUBSIDIO_ASIGNADO        16269 non-null  float64
 13  ORIGEN_DE_LA_IES         16269 non-null  object
 14  IES                      16269 non-null  object
 15  DEPARTAMENTO_IES         16269 non-null  object
 16  CIUDAD_IES               16269 non-null  object
 17  PROGRAMA                 16269 non-null  object
 18  ESTADO_DEL_BENEFICIARIO  16269 non-null  int64
dtypes: float64(3), int64(4), object(12)
memory usage: 2.4+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D22**

*Matriz de confusión escenario 6*




Fuente; Elaboración propia, (2023)

**Figura D23**

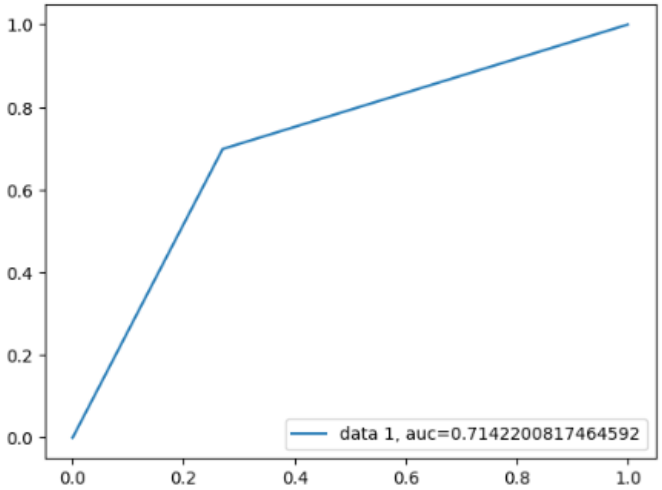
*Medidas escenario 6*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.70 | 0.73 | 0.71 | 1628 |
| 1 | 0.73 | 0.70 | 0.71 | 1706 |
| accuracy | | | 0.71 | 3334 |
| macro avg | 0.71 | 0.71 | 0.71 | 3334 |
| weighted avg | 0.71 | 0.71 | 0.71 | 3334 |

Fuente; Elaboración propia, (2023)

**Figura D24**

*Curva ROC escenario 6*




Fuente; Elaboración propia, (2023)

**Figura D25**

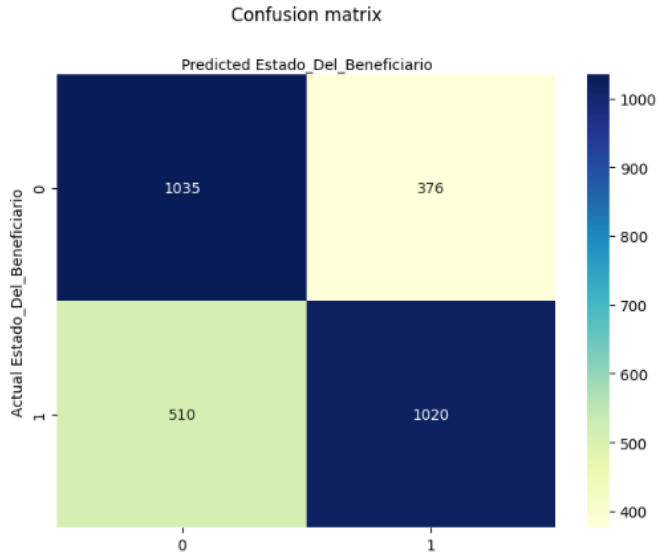
*Variables escenario 7*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 24 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   No._Beneficiario          16269 non-null  int64
 1   VERSION                   16269 non-null  object
 2   ESTRATO                   16269 non-null  int64
 3   EDAD                      16269 non-null  int64
 4   GENERO                    16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA   16269 non-null  object
 6   CIUDAD_DE_FAMILIA         16269 non-null  object
 7   EDUCACION_PADRE           14300 non-null  object
 8   EDUCACION_MADRE           14300 non-null  object
 9   OCUPACION_PADRE           14299 non-null  object
 10  OCUPACION_MADRE           14297 non-null  object
 11  DEPARTAMENTO_DE_COLEGIO   16269 non-null  object
 12  CIUDAD_DE_COLEGIO         16269 non-null  object
 13  COLE_NATURALEZA           14293 non-null  object
 14  COLE_JORNADA              14293 non-null  object
 15  PUNTAJE_SABER_11          16268 non-null  float64
 16  PUNTAJE_SISBENII          16190 non-null  float64
 17  AREA_SISBENIII            16190 non-null  object
 18  ORIGEN_DE_LA_IES          16269 non-null  object
 19  IES                       16269 non-null  object
 20  DEPARTAMENTO_IES          16269 non-null  object
 21  CIUDAD_IES                16269 non-null  object
 22  PROGRAMA                  16269 non-null  object
 23  ESTADO_DEL_BENEFICIARIO   16269 non-null  int64
dtypes: float64(2), int64(4), object(18)
memory usage: 3.0+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D26**

*Matriz de confusión escenario 7*




Fuente; Elaboración propia, (2023)

**Figura D27**

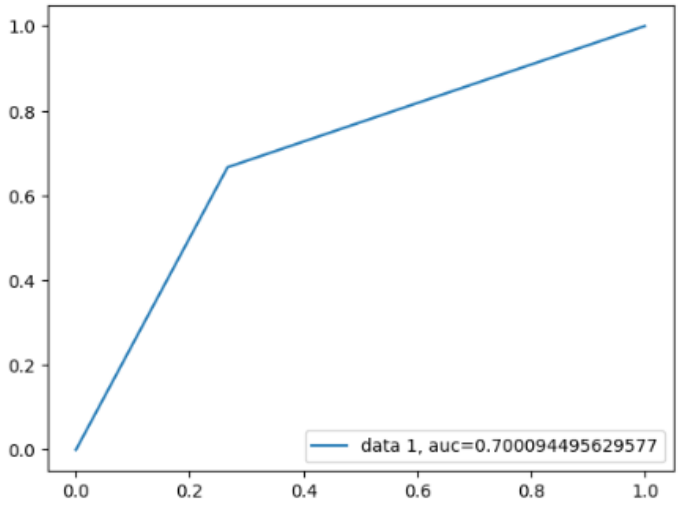
*Medidas escenario 7*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.73 | 0.70 | 1411 |
| 1 | 0.73 | 0.67 | 0.70 | 1530 |
| accuracy | | | 0.70 | 2941 |
| macro avg | 0.70 | 0.70 | 0.70 | 2941 |
| weighted avg | 0.70 | 0.70 | 0.70 | 2941 |

Fuente; Elaboración propia, (2023)

**Figura A28**

*Curva ROC escenario 7*




Fuente; Elaboración propia, (2023)

**Figura D29**

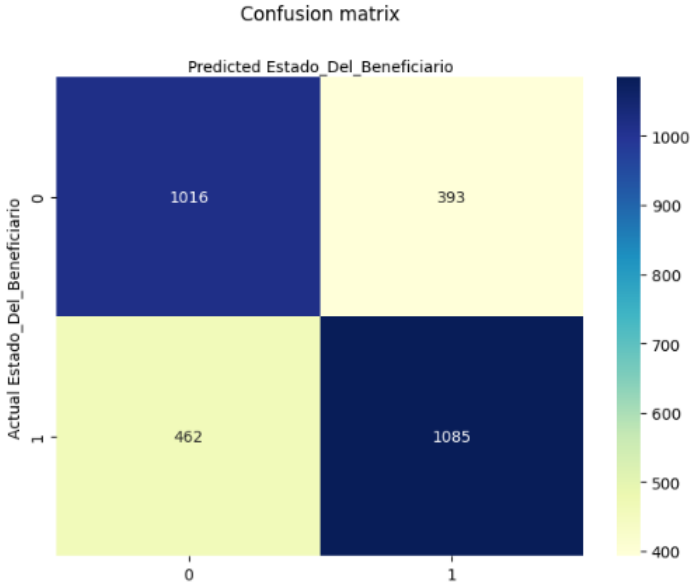
*Variables escenario 8*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 20 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   No._Beneficiario         16269 non-null  int64
 1   VERSION                  16269 non-null  object
 2   ESTRATO                  16269 non-null  int64
 3   EDAD                     16269 non-null  int64
 4   GENERO                   16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA  16269 non-null  object
 6   CIUDAD_DE_FAMILIA        16269 non-null  object
 7   DEPARTAMENTO_DE_COLEGIO  16269 non-null  object
 8   CIUDAD_DE_COLEGIO        16269 non-null  object
 9   COLE_NATURALEZA          14293 non-null  object
 10  COLE_JORNADA             14293 non-null  object
 11  PUNTAJE_SABER_11         16268 non-null  float64
 12  PUNTAJE_SISBENII         16190 non-null  float64
 13  AREA_SISBENIII           16190 non-null  object
 14  SUBSIDIO_ASIGNADO        16269 non-null  float64
 15  ORIGEN_DE_LA_IES         16269 non-null  object
 16  IES                      16269 non-null  object
 17  DEPARTAMENTO_IES         16269 non-null  object
 18  PROGRAMA                 16269 non-null  object
 19  ESTADO_DEL_BENEFICIARIO  16269 non-null  int64
dtypes: float64(3), int64(4), object(13)
memory usage: 2.5+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D30**

*Matriz de confusión escenario 8*




Fuente; Elaboración propia, (2023)

**Figura D31**

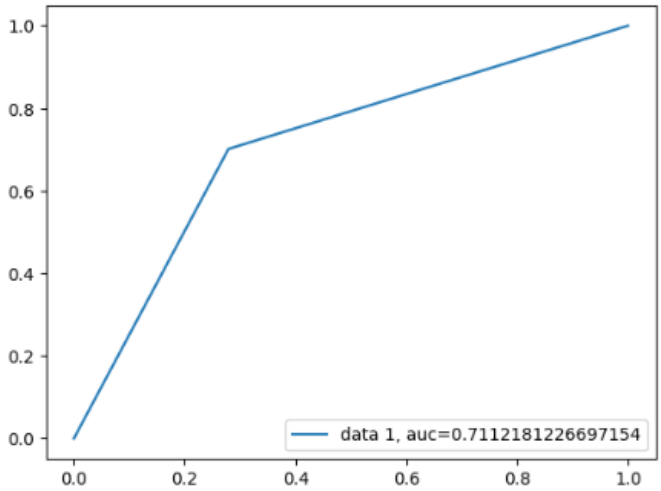
*Medidas escenario 8*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.69 | 0.72 | 0.70 | 1409 |
| 1 | 0.73 | 0.70 | 0.72 | 1547 |
| accuracy | | | 0.71 | 2956 |
| macro avg | 0.71 | 0.71 | 0.71 | 2956 |
| weighted avg | 0.71 | 0.71 | 0.71 | 2956 |

Fuente; Elaboración propia, (2023)

**Figura D32**

*Curva Roc escenario 8*




Fuente; Elaboración propia, (2023)

**Figura D33**

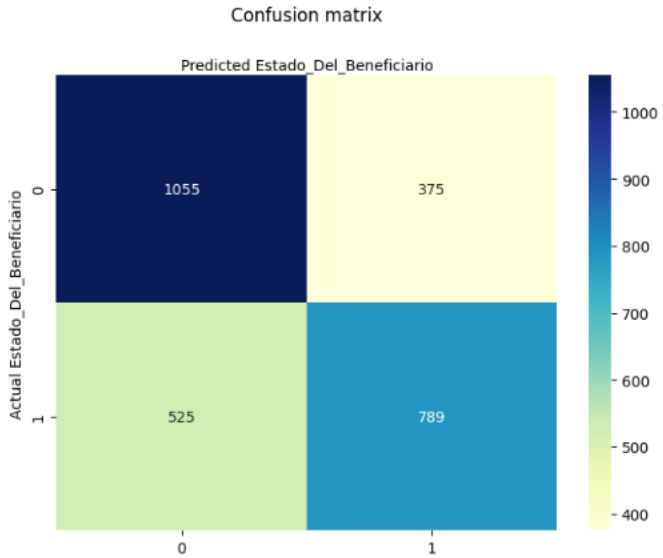
*Variables escenario 9*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 26 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   No._Beneficiario          16269 non-null  int64
 1   VERSION                   16269 non-null  object
 2   ESTRATO                   16269 non-null  int64
 3   EDAD                      16269 non-null  int64
 4   GENERO                    16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA   16269 non-null  object
 6   CIUDAD_DE_FAMILIA         16269 non-null  object
 7   EDUCACION_PADRE           14300 non-null  object
 8   EDUCACION_MADRE           14300 non-null  object
 9   OCUPACION_PADRE           14299 non-null  object
 10  OCUPACION_MADRE           14297 non-null  object
 11  INGRESO_FAMILIAR_MENSUAL  13794 non-null  object
 12  DEPARTAMENTO_DE_COLEGIO   16269 non-null  object
 13  CIUDAD_DE_COLEGIO         16269 non-null  object
 14  COLE_NATURALEZA           14293 non-null  object
 15  VALOR_PENSION_COLEGIO     13745 non-null  object
 16  COLE_JORNADA              14293 non-null  object
 17  PUNTAJE_SABER_11          16268 non-null  float64
 18  PUNTAJE_SISBENII          16190 non-null  float64
 19  AREA_SISBENIII            16190 non-null  object
 20  SUBSIDIO_ASIGNADO         16269 non-null  float64
 21  ORIGEN_DE_LA_IES          16269 non-null  object
 22  IES                       16269 non-null  object
 23  DEPARTAMENTO_IES          16269 non-null  object
 24  CIUDAD_IES                16269 non-null  object
 25  ESTADO_DEL_BENEFICIARIO   16269 non-null  int64
dtypes: float64(3), int64(4), object(19)
memory usage: 3.2+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D34**

*Matriz de confusión escenario 9*




Fuente; Elaboración propia, (2023)

**Figura D35**

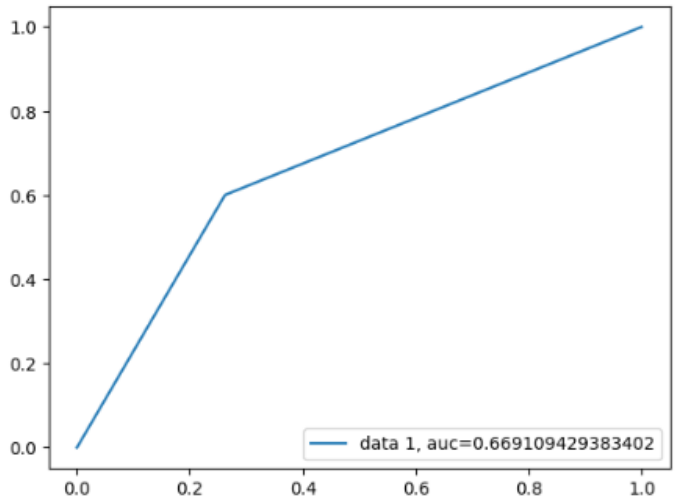
*Medidas escenario 9*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.74 | 0.70 | 1430 |
| 1 | 0.68 | 0.60 | 0.64 | 1314 |
| accuracy | | | 0.67 | 2744 |
| macro avg | 0.67 | 0.67 | 0.67 | 2744 |
| weighted avg | 0.67 | 0.67 | 0.67 | 2744 |

Fuente; Elaboración propia, (2023)

**Figura D36**

*Curva ROC escenario 10*




Fuente; Elaboración propia, (2023)

**Figura D37**

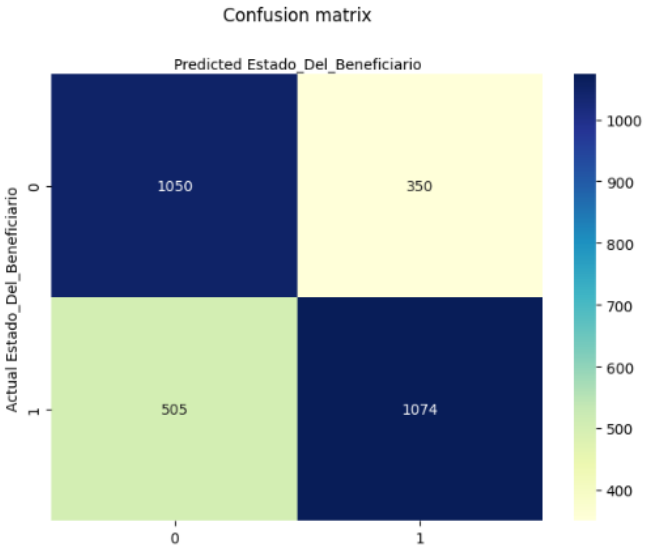
*Variables Escenario 10*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   No._Beneficiario          16269 non-null  int64
 1   VERSION                   16269 non-null  object
 2   ESTRATO                   16269 non-null  int64
 3   EDAD                      16269 non-null  int64
 4   GENERO                    16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA   16269 non-null  object
 6   CIUDAD_DE_FAMILIA         16269 non-null  object
 7   EDUCACION_PADRE           14300 non-null  object
 8   EDUCACION_MADRE           14300 non-null  object
 9   OCUPACION_PADRE           14299 non-null  object
 10  OCUPACION_MADRE           14297 non-null  object
 11  DEPARTAMENTO_DE_COLEGIO   16269 non-null  object
 12  CIUDAD_DE_COLEGIO         16269 non-null  object
 13  COLE_NATURALEZA           14293 non-null  object
 14  COLE_JORNADA              14293 non-null  object
 15  PUNTAJE_SABER_11          16268 non-null  float64
 16  SUBSIDIO_ASIGNADO         16269 non-null  float64
 17  ORIGEN_DE_LA_IES          16269 non-null  object
 18  IES                       16269 non-null  object
 19  DEPARTAMENTO_IES          16269 non-null  object
 20  CIUDAD_IES                16269 non-null  object
 21  PROGRAMA                  16269 non-null  object
 22  ESTADO_DEL_BENEFICIARIO   16269 non-null  int64
dtypes: float64(2), int64(4), object(17)
memory usage: 2.9+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D38**

*Matriz de confusión escenario 10*




Fuente; Elaboración propia, (2023)

**Figura D39**

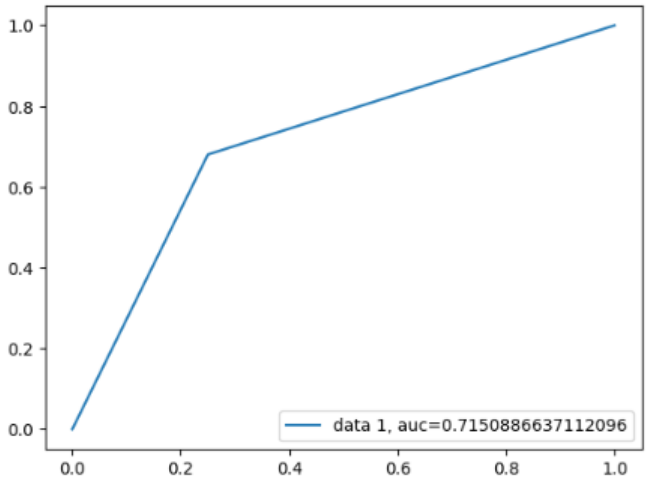
*Medidas escenario 10*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.75 | 0.71 | 1400 |
| 1 | 0.75 | 0.68 | 0.72 | 1579 |
| accuracy | | | 0.71 | 2979 |
| macro avg | 0.71 | 0.72 | 0.71 | 2979 |
| weighted avg | 0.72 | 0.71 | 0.71 | 2979 |

Fuente; Elaboración propia, (2023)

**Figura D40**

*Curva ROC escenario 10*




Fuente; Elaboración propia, (2023)

**Figura D41**

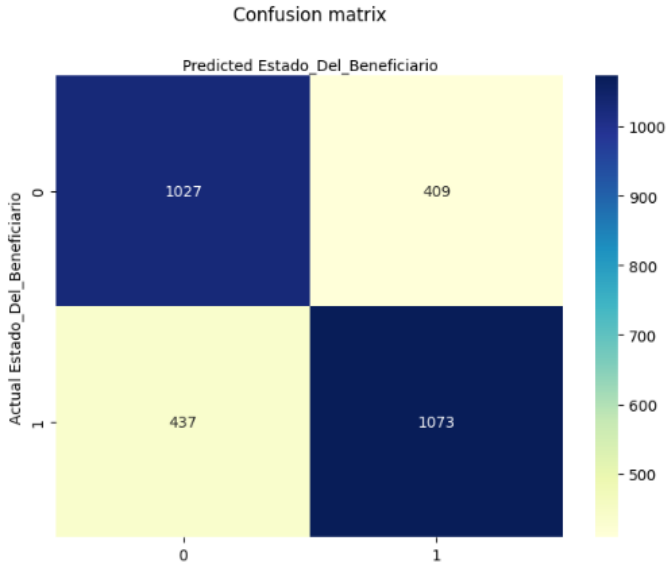
*Variables escenario 11*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 19 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   No._Beneficiario          16269 non-null  int64
 1   VERSION                   16269 non-null  object
 2   ESTRATO                   16269 non-null  int64
 3   EDAD                      16269 non-null  int64
 4   GENERO                    16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA   16269 non-null  object
 6   CIUDAD_DE_FAMILIA         16269 non-null  object
 7   DEPARTAMENTO_DE_COLEGIO   16269 non-null  object
 8   CIUDAD_DE_COLEGIO         16269 non-null  object
 9   COLE_NATURALEZA           14293 non-null  object
 10  COLE_JORNADA              14293 non-null  object
 11  PUNTAJE_SABER_11          16268 non-null  float64
 12  SUBSIDIO_ASIGNADO         16269 non-null  float64
 13  ORIGEN_DE_LA_IES          16269 non-null  object
 14  IES                       16269 non-null  object
 15  DEPARTAMENTO_IES          16269 non-null  object
 16  CIUDAD_IES                16269 non-null  object
 17  PROGRAMA                  16269 non-null  object
 18  ESTADO_DEL_BENEFICIARIO   16269 non-null  int64
dtypes: float64(2), int64(4), object(13)
memory usage: 2.4+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D42**

*Matriz de confusión escenario 11*




Fuente; Elaboración propia, (2023)

**Figura D43**

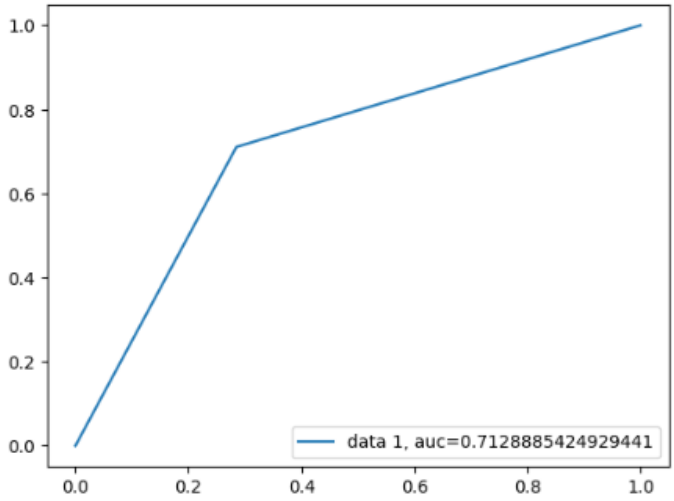
*Medidas escenario 11*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.70 | 0.72 | 0.71 | 1436 |
| 1 | 0.72 | 0.71 | 0.72 | 1510 |
| accuracy | | | 0.71 | 2946 |
| macro avg | 0.71 | 0.71 | 0.71 | 2946 |
| weighted avg | 0.71 | 0.71 | 0.71 | 2946 |

Fuente; Elaboración propia, (2023)

**Figura D44**

*Curva ROC escenario 11*




Fuente; Elaboración propia, (2023)

**Figura D45**

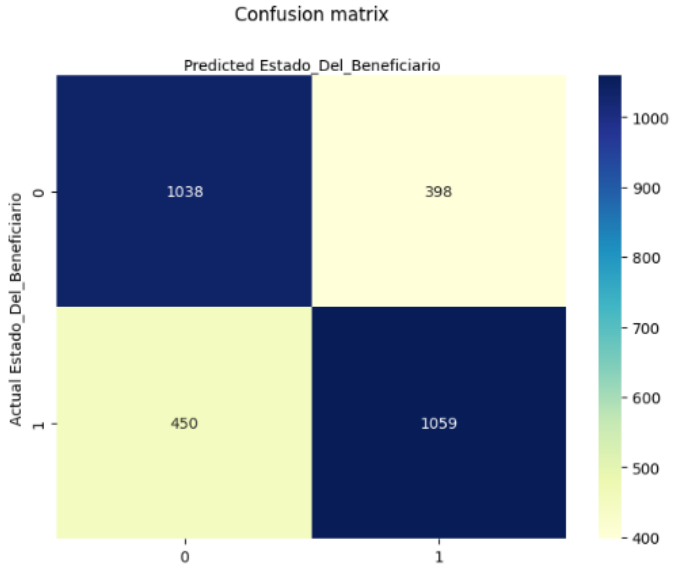
*Variables escenario 12*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16269 entries, 0 to 16268
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   No._Beneficiario         16269 non-null  int64
 1   VERSION                  16269 non-null  object
 2   ESTRATO                  16269 non-null  int64
 3   EDAD                     16269 non-null  int64
 4   GENERO                   16269 non-null  object
 5   DEPARTAMENTO_DE_FAMILIA  16269 non-null  object
 6   CIUDAD_DE_FAMILIA        16269 non-null  object
 7   DEPARTAMENTO_DE_COLEGIO  16269 non-null  object
 8   CIUDAD_DE_COLEGIO        16269 non-null  object
 9   COLE_NATURALEZA          14293 non-null  object
 10  COLE_JORNADA             14293 non-null  object
 11  PUNTAJE_SABER_11         16268 non-null  float64
 12  SUBSIDIO_ASIGNADO        16269 non-null  float64
 13  ORIGEN_DE_LA_IES         16269 non-null  object
 14  IES                      16269 non-null  object
 15  DEPARTAMENTO_IES         16269 non-null  object
 16  PROGRAMA                 16269 non-null  object
 17  ESTADO_DEL_BENEFICIARIO  16269 non-null  int64
dtypes: float64(2), int64(4), object(12)
memory usage: 2.2+ MB
```

Fuente; Elaboración propia, (2023)

**Figura D46**

*Matriz de confusión escenario 12*




Fuente; Elaboración propia, (2023)

**Figura D47**

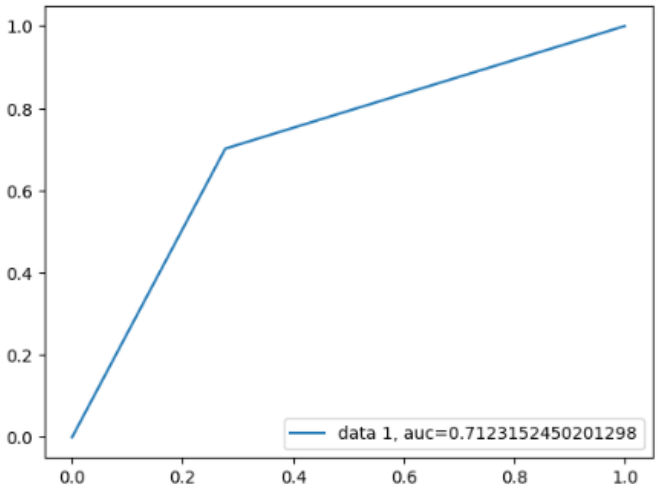
*Medidas escenario 12*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.70 | 0.72 | 0.71 | 1436 |
| 1 | 0.73 | 0.70 | 0.71 | 1509 |
| accuracy | | | 0.71 | 2945 |
| macro avg | 0.71 | 0.71 | 0.71 | 2945 |
| weighted avg | 0.71 | 0.71 | 0.71 | 2945 |

Fuente; Elaboración propia, (2023)

**Figura D48**

*Curva ROC escenario 12*




Fuente; Elaboración propia, (2023)